Universidad Autónoma de Madrid

Departamento de Biología Molecular

Facultad de Ciencias

Hacia una Visión más Dinámica de la Bioinformática Estructural en el Diseño de Fármacos y Avances en la Terapia Personalizada: Desarrollo y Aplicaciones

Helena Isabel Gomes Dos Santos

MADRID 2013

Departamento de Biología Molecular
Facultad de Ciencias
Universidad Autónoma de Madrid

Hacia una Visión más Dinámica de la Bioinformática Estructural en el Diseño de Fármacos y Avances en la Terapia Personalizada: Desarrollo y Aplicaciones

Memoria presentada por
Helena Isabel Gomes Dos Santos

Directores de Tesis
Dr. Antonio Jesús Morreale de León y Dr. Ugo Bastolla Bufalini

Tutora
Dra. Beatriz López Corcuera

Este trabajo ha sido realizado en la Unidad de Bioinformática del Centro de Biología Molecular Severo Ochoa, UB-CBMSO (Departamento de Biología Molecular).

A mi familia y a Carlos

Agradecimientos

En primer lugar quiero agradecer la oportunidad que hace unos años me brindaron desde la Unidad de Bioinformática para entrar a formar parte de su equipo. Un grupo que desde sus comienzos, con Ángel Ramírez Ortiz, hasta nuestros días ha cosechado numerosos éxitos reflejados en sus publicaciones, que van desde el desarrollo de metodologías computacionales hasta su aplicación en un contexto biológico. Estoy orgullosa el haber aprendido y trabajado en un entorno científico tan favorable y enriquecedor. Un equipo que, a pesar de las dificultades financieras por las que ha pasado la Unidad, entre otras, no ha perdido su motivación por sacar adelante los proyectos en curso y por seguir generando conocimiento relevante para el bienestar de las personas (por ejemplo a través de su línea de desarrollo de fármacos).

Quiero reconocer la labor de Ugo y de Antonio en su lucha diaria durante estos últimos años, no solo por los retos a los que se enfrentan cada día en su campo de investigación, sino por su búsqueda de financiación y estabilidad que permitieron sacar el grupo adelante tras la pérdida de Ángel. Quiero agradecer además a todos los miembros del laboratorio, tanto los que siguen aquí como lo que tuvieron que marcharse, por haberme acogido y ayudado desde el principio cada vez que lo he necesitado. Por otra parte quiero agradecer a Beatriz su motivación y buena disposición constantes, tan necesarias sobre todo en estos últimos meses. A Antonio, Ugo y Beatriz quiero agradecerles especialmente el tiempo y el conocimiento que me han aportado a lo largo de estos años, ayudándome a crecer como científica y como persona.

Agradezco además la oportunidad que he tenido de haber trabajado en colaboración con numerosos grupos experimentales y en proyectos tan diversos, lo que me ha permitido aprender un sinfín de metodologías computacionales y aplicarlas a casos reales.

Resumen

El diseño de nuevos fármacos surge de las necesidades de un entorno cambiante, donde la aparición de nuevas enfermedades y de variantes que provocan resistencia a los fármacos existentes supone nuevos retos que las técnicas experimentales y computacionales en conjunto deben resolver, manteniéndose, para ello, en constante actualización. Al mismo tiempo, nuestro conocimiento acerca de la estructura, la energética y la dinámica de las proteínas, así como las interacciones que establecen entre sí y con pequeñas moléculas está avanzando enormemente. Ello nos proporciona la oportunidad de desarrollar estrategias de cálculo más realistas durante el diseño racional de fármacos basado en estructura, que pueden explorar una parte importante del espacio químico en un tiempo razonable. Tales estrategias son nuestro principal objetivo.

En esta tesis se presentan métodos y protocolos de bioinformática estructural orientados al diseño de fármacos y sus aplicaciones a dianas de interés biomédico. Los principales resultados y conclusiones de los trabajos de investigación presentados aquí se recogen en 7 artículos (6 publicaciones y 1 manuscrito en preparación), fruto de la colaboración con diversos grupos experimentales. Los métodos que se han estudiado incluyen: **(1)** Un protocolo automático de modelado por homología combinado con una base de datos de perfiles de estructuras resueltas experimentalmente que ha permitido la generación y el refinado de numerosas estructuras 3D de dianas terapéuticas, **(2)** las mejoras de un protocolo de *docking* y cribado virtual con el fin de aumentar su capacidad predictiva y obtener candidatos más plausibles, **(3)** el estudio de la flexibilidad de proteínas tanto de modo masivo (validación de un modelo nulo de la relación entre los cambios de conformación de las proteínas y sus modos normales torsionales) como a nivel individual (dinámica molecular de un conjunto de péptidos presentados por la proteína HLA-B27*05 del complejo MHC-I en humanos) y **(4)** el desarrollo de un protocolo de control de calidad y análisis de datos provenientes de *microarrays* para la identificación de mutaciones implicadas en la resistencia a fármacos conocidos. Estos métodos se han aplicado a proyectos de investigación biomédica en campos como la biología molecular, inmunología, virología y microbiología. Los sistemas biológicos estudiados incluyen: **(a)** las proteínas que conforman el centrosoma humano, **(b)** el complejo mayor de histocompatibilidad (MHC-I), **(c)** la β-lactamasa bacteriana OXA-24 y **(d)** la proteasa (PR) y la retro-transcriptasa (RT) del virus VIH. Todas ellas de interés por ser dianas terapéuticas en enfermedades humanas con relevancia para el desarrollo de nuevos fármacos. A su vez, la correcta identificación de las mutaciones de resistencia a fármacos presentes en cada paciente nos servirá de guía en la aplicación de la medicina personalizada en los próximos años.

Abstract

Drug design arises from the requirements of a continuously changing environment where new diseases and resistance variants to existing drugs introduce new challenges that both experimental and computational techniques must resolve together, being constantly updated. At the same time, our knowledge of the structure, energetics and dynamics of proteins and their interactions with other proteins and with small molecules is greatly improving. This provides the opportunity for developing more informed computational strategies for structure-based rational drug design that can explore a huge chemical space in a reasonable time. Such strategies are our main objective.

This thesis presents methods and protocols in structural bioinformatics focused on the drug design process and their applications to interesting biomedical targets. The results and conclusions of the research projects presented here are collected in 7 papers (6 publications and one manuscript in preparation), resulting from a number of collaborations with experimental groups. Studied methods include: **(1)** An automated homology modelling protocol combined with a profile's database of experimentally solved structures, which has allowed the generation and further refinement of numerous 3D structures for therapeutically relevant targets, **(2)** the improvement of the docking and virtual screening protocols to increase their accuracy and to obtain more plausible hits, **(3)** the study of protein flexibility at both massively (validation of a null model of the relationship between the conformational changes of proteins and their torsional normal modes) and for individual proteins (molecular dynamics simulations of a peptides set which bind to HLA-B27*05, a protein within the human MHC-I complex) and **(4)** the development of a data analysis and quality control protocol of hybridization signals from microarrays, in order to identify mutations involved in known drugs resistance. These methods have been applied to biomedical research projects in fields such as molecular biology, immunology, virology and microbiology. Biological systems studied include: **(a)** proteins comprising the human centrosome, **(b)** the major histocompatibility complex (MHC- I), **(c)** the bacterial β-lactamase OXA- 24 and **(d)** the protease (PR) and retro-transcriptase (RT) HIV proteins. All of them are of interest as therapeutic targets in human diseases with relevance in the development of new drugs. Moreover, the correct identification of drug resistance mutations presented in each patient will guide personalized medicine in the upcoming years.

Índice general

Índice de figuras

Abreviaturas

3D: Tridimensional

ADN: Ácido desoxirribonucleico:

ANM (*Anisotropic Network Model*): Modelo de red anisotrópico

CBMSO: Centro de Biología Molecular Severo Ochoa

CSIC: Consejo Superior de Investigaciones Científicas

CTL(*Citolytic T Lymphocyte*):linfocitos T citotóxicos

Curva ROC (*Receiver-operating characteristic plot*): Curva característica operativa del receptor

DOF (*Degree Of Freedom*): Grados de libertad

ED (*Essential Dynamics*): Dinámica esencial

ENM (*Elastic Network Model*): Modelo de red elástica

FN (*False Negative*): Falso negativo

FP (*False Positive*): Falso positivo

GB: Generalizado de Born

GDT (*Global Distance Test*): Test de distancia global

GPU (*Graphics Processing Unit*): unidad de procesamiento gráfico

HB (*Hydrogen Bond*): Enlace de hidrógeno

HLA (*Human Leukocyte Antigen*): Complejo mayor de histocompatibilidad en humanos

HM (*Homology modeling*): Modelado por homología o modelado molecular

ISM (*Implicit Solvent Model*): Modelo de solvente implícito

LE (Ligand Efficiency): Eficacia del ligando

MD (*Molecular Dynamics*): Dinámica molecular

MHC (*Major Histocompatibility Complex*): Complejo mayor de histocompatibilidad

MM (*Molecular Mechanics*): Mecánica molecular

NMA (*Normal Mode Analysis*): Análisis de los modos normales de vibración

ns: nanosegundo

PB: Poisson-Boltzmann

PCA (*Principal Component Analysis*): Análisis de componentes principales

PCM (*Pericentriolar material*): Material pericentriolar

PR: Proteasa

RMN: Resonancia magnética nuclear

RMSD (*Root Mean Square Deviation*): Desviación cuadrática media

RT: Retrotranscriptasa

RX: Cristalografía de rayos X

SA (*Surface Area*): Area superficial

SCS (*Structurally Constrained Substitutions*): Sustituciones con restricciones estructurales

SMILES (*Simplified Molecular Input Line Entry Specification*): Estándar para la representación de estructuras moleculares usando cadenas de una dimensión

SNP (*Single Nucleotide Polymorphism*): polimorfismos de nucleótido individual

TCR (*T cell receptor*): Receptor de linfocitos T

TN (*True Negative*): Verdadero negativo

TP (*True Positive*): Verdadero positivo

QM (*Quantum Mechanics*): Mecánica cuántica

vdW: van der Waals

VIH (*HIV*, *Human Immunodeficiency Virus*): Virus de la inmunodeficiencia adquirida

VS (*Virtual Screening*): Cribado virtual

VSDMIP (*Virtual Screening Data Management on an Integrated Platform*): Plataforma para el manejo de datos provenientes de cribados virtuales

TCR (*T Cell Receptor*): Receptor de linfocitos T

WT (*Wild Type*): tipo silvestre

Motivación

La motivación principal de esta tesis parte de la necesidad de incorporar mejoras en la capacidad predictiva de las diferentes metodologías computacionales utilizadas durante el diseño de fármacos asistido por ordenador para su posterior aplicación a casos de relevancia biomédica.

Un fármaco es una molécula pequeña con un peso molecular inferior a 500 Dalton que actúa mediante interacciones moleculares con macromoléculas biológicas, como proteínas o ácidos nucleicos, modulando su actividad (Figura 1). El lanzamiento de nuevos fármacos al mercado requiere de un enorme esfuerzo en investigación, desarrollo e inversión económica. Las últimas estimaciones sitúan en casi un billón de dólares (DiMasi, Hansen, & Grabowski, 2003) y alrededor de 15 años (Myers & Baker, 2001) el dinero y el tiempo necesarios para que una molécula esté accesible al público desde su descubrimiento. Existen diversas técnicas experimentales como la química combinatoria y el cribado farmacológico de alto rendimiento (*high-throughput screening* o HTS) de librerías químicas (quimiotecas, vastas colecciones de compuestos químicos alguno de los cuales puede ser un fármaco potencial) que se usan habitualmente para probar su habilidad para modificar la función de una proteína diana. A pesar de la creciente inversión a lo largo de los últimos 30 años en el uso de dichas técnicas, éstas han producido un número de fármacos bastante por debajo de las expectativas iniciales. Por ello, los métodos experimentales por sí solos se muestran incapaces de cubrir la demanda actual de nuevos medicamentos.

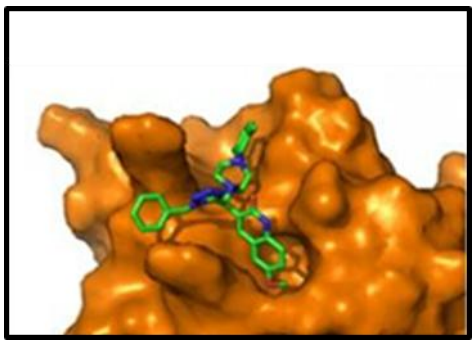


Figura 1. Ejemplo de modulación funcional mediante la unión de fármaco en el sitio activo de una proteína.

Las técnicas basadas en computación, aún con limitaciones, surgen como una buena alternativa tratando de acelerar las primeras fases del desarrollo de fármacos (Song, Lim, & Tong, 2009). El abordaje computacional engloba técnicas como el modelado por homología de estructuras tridimensionales de proteínas para los casos en los que no disponemos de ellas de modo experimental, el estudio de su comportamiento dinámico a través de modos normales y de dinámica molecular además de la predicción de sus interacciones moleculares proteína-

ligando mediante el *docking* o anclaje molecular de quimiotecas permitiendo el descubrimiento de candidatos prometedores o *hits* y su futura optimización a cabezas de serie o *leads* (Jorgensen, 2004).

Por otra parte, cada vez disponemos de un mayor número de fármacos diferentes para el tratamiento de una determinada enfermedad. En la habilidad para identificar las características genéticas propias de cada individuo, como por ejemplo mutaciones puntuales que confieran resistencia a determinados fármacos, reside nuestra capacidad para el pronóstico, diagnóstico y tratamientos clínicos eficaces. Las terapias dirigidas o personalizadas, tan prometedoras en los últimos años gracias a los últimos avances científicos y tecnológicos, nos podrían permitir la selección del mejor fármaco disponible según las características genéticas de cada individuo (Ruddy et al., 2013). Técnicas experimentales como los *microarrays* de expresión de oligonucleótidos pueden guiar en la decisión de administrar un fármaco u otro, descartando aquellos para los que un paciente presente resistencia *a priori,* evitando sus posibles efectos secundarios.

1.- Introducción

1.1.- Biología estructural

La biología es la ciencia que estudia la vida en sus diferentes niveles de organización espacial y temporal, considerando procesos macroscópicos, microscópicos, atómicos y evolutivos. Uno de los paradigmas centrales de la biología establece una relación entre la secuencia, la estructura y la función de las macromoléculas orgánicas de un organismo. Dichas moléculas son las encargadas de, mediante sus interacciones y la formación de complejos, desencadenar procesos vitales para las células (Figura 2).

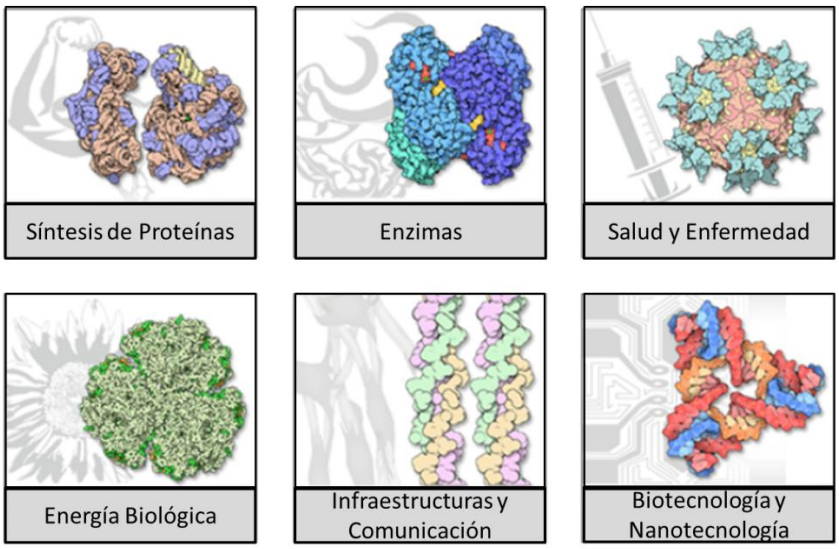


Figura 2. Estructuras de proteínas y su relevancia biológica. Ilustraciones de proteínas responsables de la síntesis de proteínas, de la catálisis enzimática, de procesos asociadas a salud y enfermedad, de la producción energética, de la generación y mantenimiento de infraestructuras, de su comunicación y regulación así como de su aplicación en el campo de la biotecnología y la nanotecnología (fuente: PDB www.rcsb.org*).*

En este marco, la biología estructural busca desentrañar el código subyacente a este paradigma aportando información experimental de proteínas y ácidos nucleicos para su caracterización estructural, flexibilidad y dinámica en condiciones fisiológicas. Las técnicas biofísicas como la cristalografía por difracción de rayos-X (RX), la cristalografía por difracción de electrones (RE), la resonancia magnética nuclear (RMN), la microscopía electrónica (ME), el análisis de partículas individuales (SP), la tomografía electrónica (TE) o la dispersión de rayos-X de bajo ángulo (SAXS), capturan información estructural detallada de los complejos macromoleculares de gran variedad de tamaños, en diversas condiciones y a distintos niveles de resolución. Sin embargo, salvo en técnicas como RNM que permiten determinar un conjunto de conformaciones representativo del estado de equilibrio termodinámico, la

información que suelen proporcionar sobre la flexibilidad y dinámica de las proteínas y ácidos nucleicos suele ser insuficiente para observar cómo se producen sus transiciones conformacionales. Debido a estas limitaciones, y gracias a los avances de las técnicas *in silico* de los últimos años, los protocolos computacionales ganan protagonismo como fuentes de información estructural.

En la presente tesis nos centraremos en la implementación y validación de nuevos métodos computacionales y su aplicación en el estudio de la estructura y la dinámica de proteínas de interés biomédico con relevancia en el desarrollo de nuevos fármacos.

En la introducción repasaremos las características químicas y físicas de los elementos que conforman las proteínas. A continuación resaltaremos el papel de la estructura y ausencia de estructura para la realización de la función proteica así como la energía asociada a su plegamiento y estabilidad. Definiremos la flexibilidad de las proteínas así como sus modelos de unión y detallaremos las interacciones moleculares que pueden establecerse entre ellas o con otras moléculas.

1.1.1.- Características de los aminoácidos

Las proteínas son polímeros lineales con funciones enzimáticas, motoras o estructurales, entre otras, donde su relevancia biológica deriva de las propiedades fisicoquímicas de los elementos que las componen: los aminoácidos. Un aminoácido, es una molécula orgánica pequeña que contiene un átomo de carbono (C_α) unido a cuatro sustituyentes: **(1)** un grupo amino (-NH_2) de naturaleza básica, **(2)** un grupo carboxilo (-COOH) de carácter ácido, **(3)** un átomo de hidrógeno y **(4)** una cadena lateral de naturaleza variable. Atendiendo a las propiedades fisicoquímicas de su cadena lateral, los 20 aminoácidos codificados genéticamente se pueden clasificar en apolares (G,A,L,V,I,M,P), aromáticos (F,Y,W), polares neutros (S,T,C,N,Q), y cargados positiva (K,H,R) o negativamente (D,E).

Los aminoácidos individualmente presentan una serie de peculiaridades que permiten determinar ciertas características y predecir la naturaleza de las proteínas ya desde su secuencia. Además, algunos aminoácidos pueden sufrir modificaciones post-traduccionales alterando la estructura y la función de la proteína. Algunas de las modificaciones post-traduccionales dirigen proteólisis limitadas permitiendo la activación o maduración de las proteínas que las incorporan, otras le aportan modificaciones químicas a los residuos existentes (fosforilación, acetilación, metilación, hidroxilación, carboxilación) influyendo en su

conformación y posibles interacciones, o incorporan otras moléculas como carbohidratos, lípidos y proteínas que pueden conferir propiedades reguladoras o funciones alternativas.

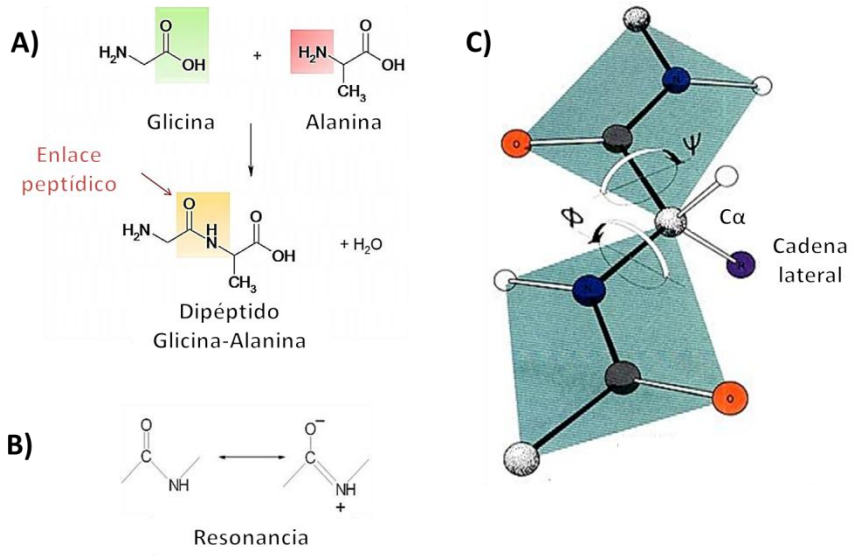


Figura 3. Enlace peptídico. **A)** Formación del enlace peptídico por condensación de los extremos C-terminal de un aminoácido y el N-terminal del siguiente, **B)** estados de resonancia del enlace peptídico, y **C)** ángulos rotables ϕ(N-Cα) y ψ (Cα-C) del esqueleto proteico o *backbone*.

La condensación de los aminoácidos individuales en péptidos y proteínas se produce mediante un enlace peptídico entre los extremos libres carboxilo, o C-terminal del primer residuo, y amino, o N-terminal del siguiente (Figura 3A). No es una reacción espontánea*, i.e.* requiere de energía para llevarse a cabo. El enlace peptídico presenta estados de resonancia (Figura 3B) entre el par de electrones del átomo de nitrógeno del esqueleto proteico y los electrones del doble enlace del grupo carbonilo C=O. La consecuencia más importante de ello es el carácter de doble enlace que se traduce en una geometría plana del enlace peptídico, determinando buena parte de las propiedades conformacionales de las proteínas. Los ángulos ϕ(N-Cα) y ψ (Cα-C) pueden rotar a lo largo del esqueleto proteico o *backbone*, mientras que el ángulo que define el enlace peptídico es casi rígido (Figura 3C).

Como veremos más adelante, a pesar de las diversas restricciones conformacionales las proteínas son moléculas con elevada flexibilidad definida tanto por los enlaces rotables (*i.e*. enlaces simples, con libertad para rotar) del *backbone* como por los grados de libertad de sus cadenas laterales.

1.1.2.- Estructura de proteínas

Hasta la segunda mitad del siglo XX, la atención de los científicos se centró en la estructura del ADN mientras que la estructura de las proteínas parecía un asunto de poca importancia, definiéndose a menudo a éstas como "*una sustancia coloidal de estructura casual*". Sin embargo, con la publicación de la primera estructura atómica completa de una proteína obtenida por RX (mioglobina) se demostró que dichas moléculas presentaban una disposición de sus átomos ordenada y específica, necesaria para ejercer su función. Desde entonces, el interés por la estructura de las proteínas así como su determinación por técnicas experimentales ha seguido un aumento exponencial (Goodsell, Burley, & Berman, 2013a). En la Figura 4 se ilustran los diferentes niveles de organización durante el plegamiento de proteínas.

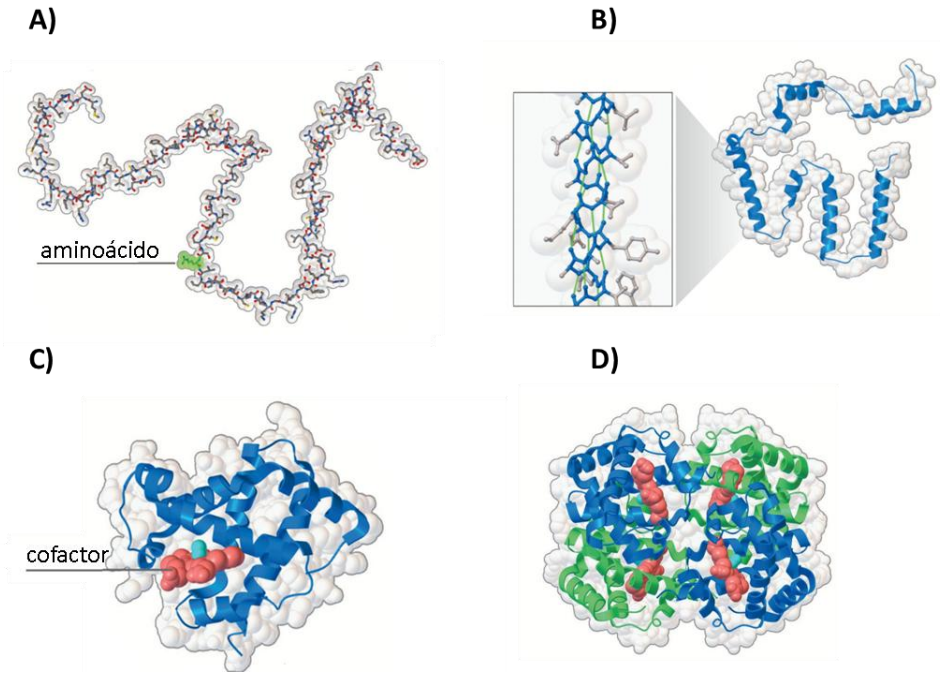


Figura 4. Plegamiento de proteínas y sus niveles de organización. A partir de la estructura primaria **(A)** o secuencia ordenada de aminoácidos, las diferentes regiones se organizan espacialmente en estructuras secundarias **(B)** promovidas por interacciones no covalentes [enlace de hidrógeno (líneas verdes) e hidrofóbicas principalmente] llamadas hélices-α, láminas-β y bucles o *loops.* El ensamblaje de dichos elementos da lugar a un plegamiento o *fold* **(C)** característico. Los dominios globulares son unidades estructurales con estabilidad independiente. Una cadena peptídica puede formar varios dominios, a veces independientes desde el punto de vista evolutivo o funcional. Cuando varias cadenas, iguales o diferentes, se unen forman estructuras cuaternarias o complejos macromoleculares **(D)** [PDB ID: 1hh0]. Además de estas estructuras principales a veces se pueden caracterizar estructuras super-secundarias de nivel intermedio entre la estructura secundaria y el dominio globular.

En el mes de marzo de 2013 el número de estructuras moleculares entre proteínas y ácidos nucleicos depositadas en el Protein Data Bank (PDB) (http://www.rcsb.org; Sussman et al., 1998), base de datos de referencia en el campo, ascendían a 88.837 (92% de proteínas globulares y de membrana), de las cuales 50.054 presentaban secuencias no redundantes, y su número mantiene una tendencia de crecimiento exponencial a lo largo del tiempo. Los principales métodos experimentales de determinación estructural en esta base de datos son RX y RNM (88% y 11%, respectivamente) representando las dos caras del comportamiento de las moléculas orgánicas: estático y dinámico, respectivamente. En el caso de RX la resolución media de los cristales es de 2.2 Å, con una desviación estándar de 1.6 Å. Sin embargo, a pesar del gran avance que ha supuesto la calidad y el volumen de datos estructurales disponibles, este número es relativamente pequeño en relación con los millones de secuencias de proteínas que se almacenan en bases de datos como UniProt (http://www.uniprot.org; Uniprot Consortium, 2013). El desfase entre el número de secuencias conocidas y el de estructuras definidas continúa creciendo a pesar de los esfuerzos de la genómica estructural por determinar dianas representativas de familias de proteínas y su plegamiento o *fold* (Chandonia & Brenner, 2006). Sin embargo, el descubrimiento de nuevos plegamientos parece no ser un factor tan limitante ya que el número de éstos en la naturaleza se estima limitado (Dokholyan, Shakhnovich, & Shakhnovich, 2002) y, como se ha venido observando, a pesar de la producción exponencial de nuevos datos la tasa de descubrimiento de nuevos plegamientos es cada vez más reducida.

Una de las observaciones empíricas en las que se basa la biología estructural computacional es que proteínas homologas (con origen evolutivo común) tienen estructuras parecidas y, aunque más limitadamente, una función similar (familia de proteínas). A día de hoy disponemos de grandes bases de datos de clasificación de familias de topologías y plegamientos [CATH, http://www.cathdb.info, (Orengo et al., 1999); y SCOP, http://scop.mrc-lmb.cam.ac.uk/scop, (Lo Conte et al., 2000)], así como de familias de funciones de proteínas (PFAM, http://pfam.sanger.ac.uk, Punta et al., 2012) y de relaciones evolutivas (OMAdb, www.omabrowser.org). Esta información es de gran utilidad para los trabajos de investigación *in silico*. La clasificación estructural de proteínas conocidas se viene realizando en base a diferentes parámetros de similitud y niveles de organización: SCOP realiza una clasificación manual jerárquica que describe las relaciones estructurales y evolutivas entre sus proteínas, CATH ofrece clasificaciones jerárquicas semi-automáticas de dominios. Otros grupos proponen que una clasificación automática basada en similitud estructural sólo puede ser internamente

consistente por similitud alta, debajo de la cual el espacio de estructuras aparece como un continuo (Pascual-García, Abia, Ortiz, & Bastolla, 2009).

1.1.2.1.- Variabilidad estructural

Existen diferentes medidas para la cuantificación de la variabilidad estructural tanto entre conformaciones de una misma molécula como para la comparación de moléculas diferentes. Minimizando esas medidas podemos obtener alineamientos estructurales entre proteínas o pequeñas moléculas que puede ser de gran ayuda para la evaluación de modelos teóricos y para la comparación de las conformaciones adoptadas por una proteína (Lemmen and Lengauer 2000). A continuación expondremos algunas de las medidas de similitud estructural más utilizadas:

(1) Desviación cuadrática media o RMSD (*root-mean-square deviation*) calcula la desviación cuadrática media entre las posiciones de los átomos de dos estructuras, *A* y *B*.

Ecuación 1: $\mathrm{RMSD}_{\mathrm{AB}} = \sqrt{\frac{1}{n}\sum_{\mathrm{i}=1}^{\mathrm{n}} \left|\left|r_i^A - T(r_i^B)\right|\right|^2}$

donde *n* es el número de átomos y $\boldsymbol{r_i^A r_i^B}$ son las coordenadas cartesianas de los átomos correspondientes a las conformaciones *A* y *B*. Sobre las coordenadas de *B* se aplica la roto-translación *T* que minimiza el RMSD (Kabsch, 1976). Durante el *docking* y el cribado virtual de pequeñas moléculas no aplicaremos las rotaciones y translaciones señaladas en este apartado.

(2) Solapamiento de contactos (*contact overlap*) mide el número de contactos entre pares de aminoácidos *i* y *j* que son comunes a dos estructuras *A* y *B*, normalizado de forma que vale uno cuando todos los contactos coinciden.

Ecuación 2: $Contact\ overlap_{AB} = \frac{\sum_{ij} C_{ij}^{(A)} C_{ij}^{(B)}}{\sqrt{\sum_{ij} C_{ij}^{(A)} \sum_{ij} C_{ij}^{(B)}}}$

donde $C_{ij}^{(A)}$ y $C_{ij}^{(B)}$ son las matrices de contactos de los residuos *i*, *j* en las estructuras *A y B* respectivamente. Establecemos que existe un contacto entre dos residuos cuando la distancia entre alguno de los átomos pesados del par de residuos se encuentra a menos de 4.5 Å. Para evitar el ruido provocado por contactos cercanos con alta probabilidad en proteínas no relacionadas se suelen considerar en el cálculo sólo aquellos contactos que están distantes en la secuencia, con $|\mathrm{i} - \mathrm{j}| > 3$. Esta medida no requiere de una superposición estructural.

Existen otras medidas estructurales como el *TM-score* (*Template Modeling score*, Zhang and Skolnick, 2004) y la divergencia de contactos (*contact divergence*, Pascual-García et al 2010), pero no entran dentro de los objetivos de esta tesis.

1.1.2.2.- Evaluación estructural

Para evaluar la calidad de estructuras determinadas experimentalmente se comparan sus propiedades con las propiedades típicas de una gran base de datos. Entre las propiedades que se evalúan estadísticamente están la distancia entre pares de átomos, los ángulos de enlace, los ángulos de torsión y la presencia de contactos entre aminoácidos observados más o menos frecuentemente.

En particular, los diagramas de *Ramachandran* representan los valores de los ángulos ϕ y ψ del esqueleto proteico. Existen valores prohibidos porque determinan repulsiones entre átomos, y valores favorecidos que corresponden a elementos de estructura secundaria como las hélices-α y las láminas-β La preferencia por determinadas regiones del diagrama depende del tipo de amino ácido considerado.

1.1.2.3.- Variabilidad genética y su repercusión en la estructura de proteínas

Las secuencias de un tipo determinado de proteína varían incluso dentro de una misma población. Un ejemplo de esta variabilidad genética son los polimorfismos de nucleótido individual (SNPs). La mayoría de los SNPs no alteran significativamente la actividad de la proteína, a veces porque están compensados por otros mecanismos celulares, pero algunos producen enfermedades. Por otra parte, virus con altas tasas de mutación, como el VIH, presentan proteínas con una variabilidad muy grande pudiendo promover la resistencia del virus tanto contra el sistema inmune del hospedador como contra los fármacos que usan la proteína mutada como diana. Otras fuentes de variabilidad genética pueden ser las duplicaciones, inserciones y delecciones.

Todas estas modificaciones a nivel de secuencia pueden repercutir en la conformación tridimensional de las proteínas que codifican, en su dinámica y en su capacidad de interaccionar con sustratos u otras proteínas durante la formación de complejos (Studer, Dessailly, & Orengo, 2013), por ello su identificación es crucial durante el desarrollo de nuevos fármacos, pero también durante el pronóstico/diagnóstico clínico, ayudando en la elección del mejor tratamiento disponible en cada caso.

1.1.3.- Flexibilidad de proteínas

La flexibilidad de las moléculas, definida por sus grados de libertad (DOFs) torsionales, se debe a que muchos conjuntos de enlaces permiten rotaciones coordinadas de una parte de la molécula respecto al resto con un coste energético muy limitado. Los grados de libertad del *backbone* y las cadenas laterales confieren a las proteínas su variabilidad estructural intrínseca, permitiendo alcanzar diferentes estados conformacionales (Figura 5), que luego se mantienen relativamente estables si son favorables energéticamente.

Asociando a cada conformación su valor de energía libre se obtiene un paisaje energético, una representación abstracta cuyos mínimos representan conformaciones relativamente estables y cuyas barreras de energía representan el coste asociado a un cambio de conformación. Estos movimientos de gran escala entre mínimos energéticos son frecuentemente modulados por la interacción con otras moléculas y pueden jugar un papel funcional o bien en la región de enlace, o bien en una región lejana (efecto alostérico), permitiendo alcanzar una conformación activa o inactiva. Bajando en la escala espacial y temporal podemos definir toda una jerarquía de movimientos moleculares: **(1)** movimientos colectivos naturales, *i.e.* movimientos coordinados intrínsecos a la estructura tridimensional de la proteína determinados por la topología de los contactos entre sus residuos; **(2)** rotaciones en cadenas laterales o *flips* y **(3)** vibraciones moleculares a nivel de átomos o grupos de átomos.

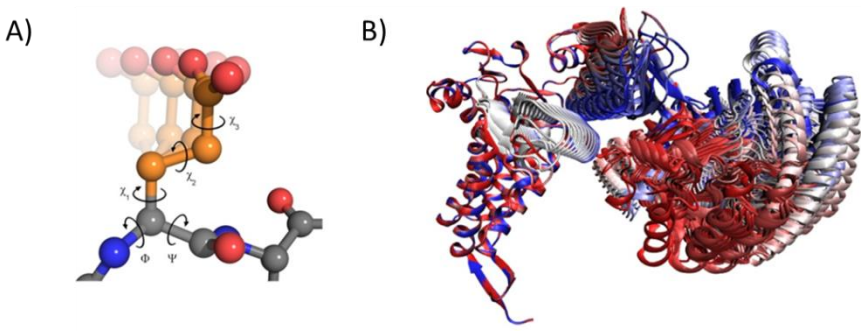


Figura 5. Ejemplos de flexibilidad de proteínas. **A)** Movilidad a nivel de residuo (ASP), **B)** Movilidad a nivel de dominio (monómero de GroEL, [PDB ID: 1SS8 y 1SX4]).

La espectroscopia de RX permite investigar las fluctuaciones térmicas de las proteínas determinando los *B-factors* o factores de temperatura de sus átomos, que miden la dispersión del átomo respecto a su posición media. Valores muy altos de los *B-factors* puede indicarnos que una región de la proteína oscila entre conformaciones alternativas. Si la flexibilidad de una región es muy alta, no es posible determinar ninguna estructura media para esta región. Este

tipo de regiones que no tienen una estructura fija se denominan regiones desordenadas. El desorden se clasifica como dinámico o dependiente de la temperatura, y estático o independiente de la temperatura si existen dos o más conformaciones estáticas alternativas.

Más adelante veremos la relación entre las diferentes conformaciones que pueden adoptar una proteína y su energía, así como su repercusión en la estabilización de algunas de ellas para su correcto funcionamiento.

1.1.4.- Proteínas intrínsecamente desordenadas

La afirmación de que una proteína requiere de una determinada organización estructural para realizar su función no quiere decir, necesariamente, que todas las proteínas de un proteoma posean una estructura definida a lo largo del tiempo. Se han evidenciado un gran número de proteínas intrínsecamente desordenadas, o IDPs, que contienen regiones de extensión variable, o incluso proteínas enteras, privadas de estructura secundaria o terciaria bajo condiciones fisiológicas (Figura 6).

Figura 6. Ejemplo de proteína con desorden estructural.

Estas regiones no estables estructuralmente se corresponden a las partes de la molécula que resultan invisibles en los mapas de densidad electrónica producidas por RX. El dicroísmo circular (DC) nos aporta evidencias experimentales de la presencia/ausencia de estructura secundaria. Mediante estas técnicas se han evidenciado transiciones desorden-orden moduladas por la interacción con otras moléculas (Uversky, 2013a).

Se ha sugerido que la flexibilidad de las IDPs tienen una importancia funcional a diferentes niveles: **(1)** la flexibilidad de una proteína puede favorecer una extensa interfaz de reconocimiento molecular, importante durante la regulación del ciclo celular, la transcripción, la traducción u otras actividades, así como el ensamblaje de otras proteínas. Las proteínas flexibles pueden unir diversos sustratos con alta especificidad y baja afinidad (considerando el gasto energético del plegamiento previo a la unión), **(2)** dado que la vida media de las proteínas en el interior celular es breve podrían representar un mecanismo rápido de *turnover,*

i.e. un balance entre la síntesis y la degradación proteica, de importantes reguladores moleculares y **(3)** generalmente las proteínas desordenadas evolucionan más rápido que las ordenadas. Shaiu *et al.* observaron que las regiones desordenadas presentes en la topoisomerasa II mostraban un número elevado de sustituciones aminoacídicas y más inserciones y delecciones que la región ordenada de esa misma molécula (Shaiu *et al.*, 1999)

1.1.5.- Relación estructura–energía en proteínas: plegamiento y estabilidad

Comprender como se pliegan y mantienen las estructuras de las proteínas ha supuesto un enorme esfuerzo científico en las últimas cinco décadas. La visión clásica del plegamiento propone que la búsqueda del estado nativo en el inmenso espacio conformacional que la proteína explora transcurre a través de determinadas rutas definidas por intermediarios y barreras energéticas. Mediante la sinergia de experimentación y teoría ha ido emergiendo un nuevo punto de vista que propone la existencia de un continuo entre estados intermedios y trayectorias convergentes de plegamiento. La nueva visión del plegamiento establece que las proteínas no siguen una ruta prefijada con intermedios obligatorios, sino que la mera pendiente del paisaje energético, a modo de embudo entrópico (*folding funnel*) conduce su plegamiento en un pequeño intervalo de tiempo.

El diámetro del embudo representa la entropía de las proteínas y la profundidad representa la ganancia de energía libre cuando la proteína se acerca a su estado nativo. Se asume que la proteína en forma desplegada se encuentra en el borde superior del embudo, con entalpía poco óptima pero entropía favorable pudiendo moverse entre diferentes estados *quasi* degenerados energéticamente. La entropía disminuye a medida que lo hace el diámetro del embudo reduciendo el número de estados accesibles pero aumentando su estabilidad energética. Las proteínas difieren entre sí en la exploración del paisaje de energía de su plegamiento condicionado por el microambiente al que están expuestas, la secuencia precisa de aminoácidos que presentan y la topología de sus cadenas (Baker, 2000, Figura 7).

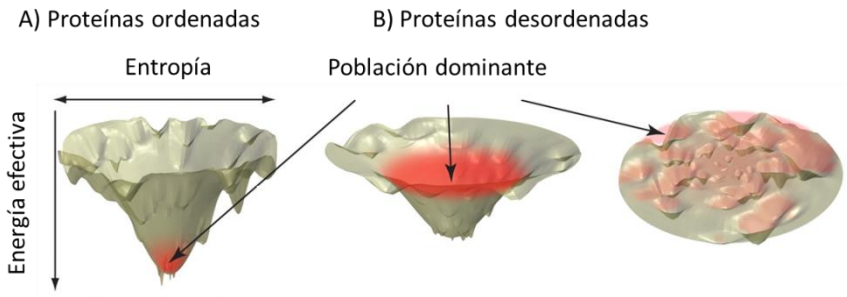


Figura 7. Ejemplo de paisajes energéticos donde la zona gris corresponde a diferentes estados, total o parcialmente desplegados, y la zona roja representan las conformaciones dominantes que se

corresponden con el estado nativo: (**A)** paisaje energético típico de proteínas ordenadas y (**B)** paisajes energéticos de proteínas con diferente grado de desorden.

El estado nativo de una proteína globular es estable en cierto intervalo de temperatura fuera del cual su estructura se desnaturaliza. La estabilidad se puede medir como *energía libre de Gibbs* de plegamiento ($\Delta G_{plegamiento}$), es decir la diferencia de energía libre entre el estado plegado y el estado desplegado. El plegamiento de una proteína conduce a mínimos de energía libre globales o locales. El proceso de plegamiento se suele representar mediante un paisaje de energía, una representación abstracta que asocia un valor de energía libre a un conjunto de conformaciones o microestados. A pesar de la enorme complejidad del espacio de conformaciones, éstas se suelen proyectar sobre un número muy reducido de direcciones. La representación más usada utiliza una sola dimensión, la coordenada de reacción (Figura 8).

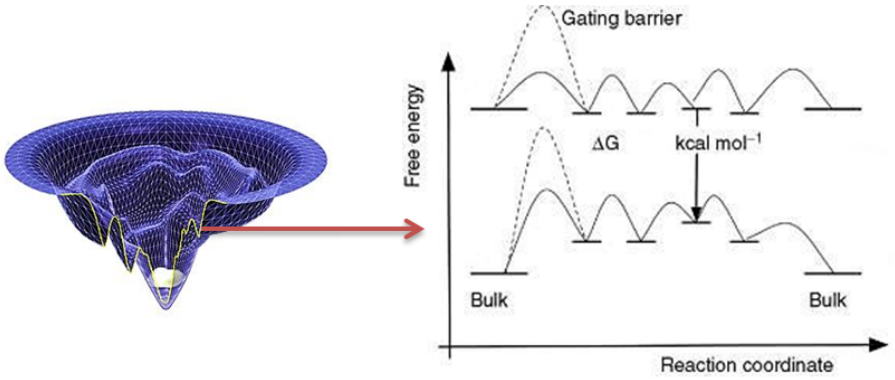


Figura 8. Energía libre de plegamiento (eje y) asociada a una coordenada de reacción o variable de progreso (eje x) y ejemplo de barreras de energía asociadas.

La variación de energía libre ($\Delta \boldsymbol{G}$) en el proceso de plegamiento vendrá dada por el balance de dos términos energéticos: el entálpico (*H*) y el entrópico (*S*), Ecuación 3, y puede ser medida experimentalmente.

Ecuación 3: $\Delta G = \Delta H - T\Delta S$

El plegamiento debe por tanto optimizar la energía libre del sistema. La contribución entálpica viene de las interacciones de los átomos de la proteína entre sí y con el solvente, enlaces de hidrógeno e interacciones de *vdW* fundamentalmente. La entropía conformacional, una medida de los microestados compatibles con un estado macroscópico dado, caracteriza el estado nativo como un pequeño conjunto de conformaciones (*ensemble*) estructuralmente próximas, mientras que el estado desplegado es constituido por un elevado número de conformaciones diferentes en equilibrio rápido. Se piensa que el estado nativo está favorecido desde el punto de vista de la entalpía y de la entropía del solvente, mientras que el estado

desplegado está favorecido desde el punto de vista de la entropía conformacional de la proteína.

Tanto el plegamiento como la asociación molecular se rigen por los mismos principios de la termodinámica (Ecuación 3). Como hemos visto, la energía libre de unión es el resultado de un balance entre la energía que se opone al plegamiento o la unión y la que lo dirige. A temperatura ambiente (25^0C) las componentes entálpica y entrópica se cancelan casi totalmente, dando valores pequeños para ΔG. Esto indica que el estado nativo de las proteínas es sólo marginalmente más estable que su estado desplegado. Sin embargo, podemos incrementar experimentalmente su ΔG mediante mutaciones puntuales. Ello lleva a pensar que la naturaleza podría haber seleccionado proteínas más estables si fuera una característica ventajosa (Sanchez-Ruiz, 1995). En esta línea se ha sugerido que la relativamente baja estabilidad puede: **(1)** ir acompañada de una mayor flexibilidad, que sería necesaria para su función, **(2)** facilitar su degradación cuando resulte conveniente, **(3)** favorecer un plegamiento rápido si sus intermedios cinéticos son poco estables y **(4)** revertir plegamientos incorrectos (trampas cinéticas) de un modo más sencillo al tener que superar barreras de energía menores. Otra posible explicación selectivamente "neutral" es que la evolución no ha seleccionado proteínas más estables porque la estabilidad alcanzada es ya suficientemente elevada para garantizar su función, y la presión selectiva para aumentarla no es suficiente para contrastar la presión natural de mutación (Taverna & Goldstein, 2002, Liberles et al., 2012).

1.2.- Interacciones moleculares

Las interacciones entre moléculas orgánicas representan el lenguaje de los sistemas biológicos mediante las cuales las macromoléculas se comunican y ejercen su función. El desarrollo de la biología molecular ha permitido evidenciar cómo muchas patologías humanas están directamente relacionadas con fallos en las interacciones entre proteínas o entre éstas y otras moléculas orgánicas que impide su correcta unión. Por ello, uno de los retos de la biología actual es el de caracterizar a nivel molecular los mecanismos de acción subyacentes a enfermedades humanas además de desarrollar y probar nuevo fármacos específicos para las dianas terapéuticas identificadas. Un conocimiento profundo de los sistemas a nivel atómico, así como la correcta caracterización de las interacciones que tienen lugar durante la unión de un receptor con su ligando, pueden ayudar a modular su actividad.

1.2.1.- Modelos de unión

A lo largo de los años se han propuesto diferentes modelos de interacción molecular. Inicialmente, prevalecía una visión estática de la interacción donde cada molécula involucrada

se comportaba como un elemento rígido (modelo llave-cerradura, postulado por Emil Fischer en 1894). Posteriormente, los modelos fueron cambiando hacia una visión más dinámica y flexible de ambas moléculas, reconociendo que muchas uniones moleculares vienen acompañadas de grandes cambios de conformación. El ajuste inducido (*induded-fit*) propuesto por Daniel Koshland en 1958 propone que la unión produce las reorganizaciones estructurales necesarias para favorecer la estabilidad del complejo. El último de los modelos propuestos, llamado selección conformacional (1999), postula que la proteína puede oscilar entre varias conformaciones, entre las cuales se encuentran tanto la conformación favorecida en ausencia de ligando como la conformación que la proteína asume en la unión. Según este modelo, el ligando se une selectivamente a esta última conformación y la estabiliza (Figura 9).

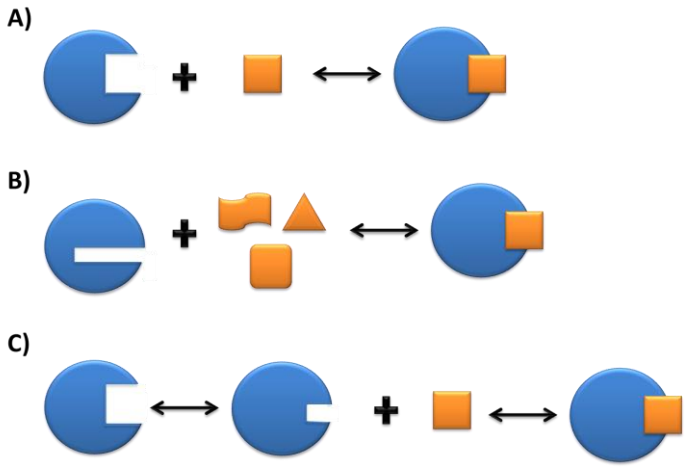


Figura 9. Esquemas de los principales modelos de unión propuestos: **A)** modelo rígido de llave-cerradura, **B)** modelo flexible de ajuste inducido y **C)** modelo flexible de selección conformacional.

En los últimos años empiezan a implantarse nuevas propuestas de modelos de unión asociados a proteínas intrínsecamente desordenadas (IDPs). Las diferentes posibilidades son: **(1)** el plegamiento inducido (*induced folding*), que representa las transiciones de desorden a orden mediante la interacción entre proteínas, **(2)** el desplegamiento funcional (*functional unfolding*), que representa la transición de orden a desorden en determinadas regiones o proteínas completas inducida por la interacción y **(3)** el modelo de unión de conformaciones imprecisas (*fuzziness*) donde la unión involucra proteínas sin plegamiento tridimensional definido, ni antes ni después de la unión (Figura 10).

La unión de pequeñas moléculas tanto a los sitios ortostéricos (sitios activos) como a los alostéricos (sitios reguladores) pueden inducir cambios conformacionales. Asociados a los sitios activos, en proteínas con estructura cuaternaria, los cambios de conformación pueden generar un comportamiento cooperativo donde la unión del primer ligando a uno de los

dominios induce cambios en las interacciones que presentan las otras subunidades, modificando la unión de las siguientes moléculas de ligando y variando la conformación global del complejo (cooperatividad positiva/negativa).

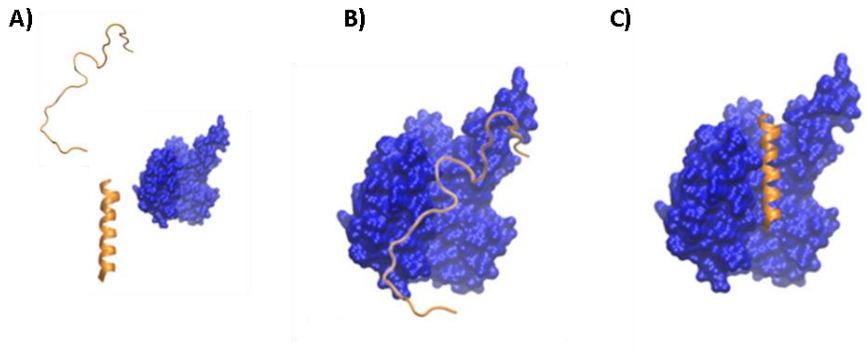


Figura 10. Modelos de unión asociados a IDPs: **A)** posibles transiciones orden-desorden, previas a la unión (por ejemplo por condiciones del microambiente), **B)** modelo de unión y estabilización de una conformación desordenada (*fuziness*) y **C)** modelo de unión y estabilización del complejo tras una transición desorden-orden (*induced folding*).

1.2.2.- Estimación de la energía libre de unión

Según la termodinámica, los modelos de unión suponen un equilibrio entre dos estados: libre y unido (*[R] + [L] ↔ [RL]* siendo *R* el receptor, *L* el ligando y *RL* el complejo formado). El tipo de interacciones involucradas en dicho equilibrio pueden ser de naturaleza enlazante o no (*i.e.* relacionadas o no con la formación de enlaces covalentes). En las siguientes secciones presentaremos las principales interacciones involucradas en la formación y estabilización de complejos moleculares.

Como vimos en la sección 1.1.5 de plegamiento y estabilidad, la *ΔG* consta de una contribución entálpica y una entrópica. Esta última se suele descartar en la estimación de $\Delta G_{unión}$ *in silico*, por su complejidad y alto coste computacional. La entalpía ($AH_{unión}$) incluye habitualmente las contribuciones de *vdW* (E_{vdw}) y electrostáticas ($E_{electrostática}$), la contribución energética asociada al establecimiento de enlaces de hidrógeno (E_{HB}) y la energía de desolvatación ($E_{desolvatación}$, Ecuación 4).

Ecuación 4: $\Delta H_{unión} = E_{electrostática} + E_{vdw} + E_{HB} + E_{desolvatación}$

1.2.3.- Interacciones enlazantes

Entre las interacciones enlazantes más características para la correcta actividad de las proteínas caben destacar lo puentes disulfuro (S-S) (Sevier & Kaiser, 2002), interacciones covalentes donde dos residuos de cisteína próximos y con la orientación adecuada se oxidan y

condensan dando lugar a un enlace covalente entre los dos átomos de azufre aportando una energía libre de unas -60 kcal/mol. Los S-S están relacionados con la estabilidad y activación de las proteínas que los contienen y su localización es normalmente extracelular, pues el potencial redox del interior de la célula es reductor y los destruye.

Otro caso de relevancia es la unión irreversible de inhibidores que conlleva la formación de enlaces covalentes entre el ligando y la proteína. La caracterización teórica de dichos enlaces sólo se puede obtener mediante metodología cuántica, no considerada en la presente tesis.

1.2.4.- Interacciones no enlazantes

A continuación definiremos las interacciones no enlazantes incluidas, en mayor o menor medida, en los modelos matemáticos de estimación de energía de unión: **(1)** las interacciones de tipo *vdW*, **(2)** interacciones electrostáticas, **(3)** interacciones apolares o hidrofóbicas, **(4)** formación de enlaces de hidrógeno, **(5)** interacciones con el solvente y **(6)** entropía.

1.2.4.1.- Interacciones de tipo van der Waals

Las interacciones de *vdW*, de gran importancia para la estabilidad de macromoléculas biológicas, cuantifican la magnitud de las fuerzas creadas entre dos átomos cuando éstos se aproximan y entran en contacto. Tienen en cuenta dos componentes: **(1)** la repulsión, que actúa a corta distancia debido al solapamiento o superposición de las nubes electrónicas de los átomos que se acercan y **(2)** la atracción que se da a larga distancia debida a la correlación entre los electrones de los diferentes átomos (fuerzas de dispersión de London). Ambas componentes dependen inversamente de la distancia entre los átomos. Las interacciones de *vdW* se describen habitualmente según el potencial 12-6 de Lennard-Jones (Figura 11, Ecuación 5).

Ecuación 5: $E_{Lennard-Jones} = 4\,\varepsilon\,\left[\left(\frac{A}{r}\right)^{12} - \left(\frac{B}{r}\right)^{6}\right]$

donde *r* es la distancia entre los átomos, *A* y *B* son los coeficientes de repulsión y atracción respectivamente que dependen del par de átomos involucrados en la interacción y ε es la profundidad del potencial.

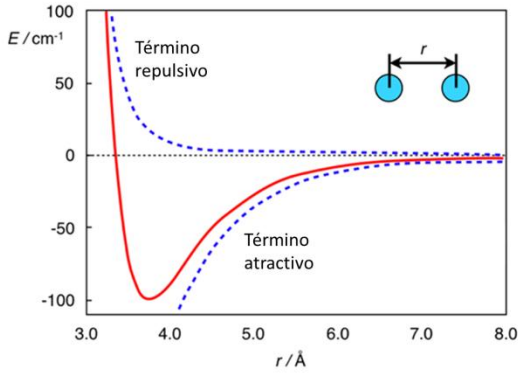


Figura 11. Potencial 12-6 de Lennard-Jones para la descripción de interacciones de tipo *vdW*.

1.2.4.2.- Interacciones electrostáticas

Estas interacciones están presentes en la mayor parte de los tipos de unión (*i.e*. interacciones carga-carga, enlaces de hidrógeno, apilamiento de nubes π o π-π *stacking*, interacciones hidrofóbicas y solvatación). Las interacciones electrostáticas (atractivas y repulsivas) de las proteínas surgen de la presencia de grupos cargados o grupos con distribuciones de carga no homogénea, como dipolos y cuadrupolos. Por otra parte, el agua ejerce un potente efecto modificador de estas interacciones, apantallándolas de modo que la atracción entre cargas contrarias es relativamente débil en la superficie expuesta al solvente. Estas fuerzas presentan una alta selectividad, un aspecto clave en el desarrollo de nuevos, y más específicos fármacos que no produzcan efectos secundarios. Sin embargo, su cálculo exacto sigue siendo uno de los mayores retos dela biología computacional. La aproximación más simple es el modelo Coulombico (E_{coul}, el producto de las cargas dividido por la distancia y una función simple para la constante dieléctrica que refleje las propiedades de respuesta del medio, Ecuación 6). La precisión del cálculo dependerá de una correcta asignación de cargas.

Ecuación 6: $E_{coul} = \frac{1}{\varepsilon} \frac{q_1 q_2}{r}$

donde ε es la constante dieléctrica del medio, q_1 y q_2 las cargas de los átomos involucrados y *r* la distancia entre ellos.

1.2.4.3.- Efecto del solvente

Muchas de las interacciones a tener en cuenta en biología tienen lugar en un entorno acuoso. Cuando las moléculas están aisladas en disolución, están completamente rodeadas de moléculas de agua. Sin embargo cuando se produce una interacción muchas de estas moléculas de agua son desplazadas. Este desplazamiento o desolvatación conlleva un gasto energético que debe ser contrarrestado por las nuevas interacciones formadas. Además, se

produce una ganancia de entropía en las moléculas de agua liberadas. Desde el punto de vista teórico hay dos modelos extremos para tener en cuenta los efectos del solvente que se aplican según el compromiso entre precisión y rapidez que necesitemos en cada caso: **(1)** modelos de solvente explícito, donde las moléculas de agua están explícitamente representadas con detalle atómico y **(2)** modelos de solvente implícito**,** donde se construye una función matemática que modela el comportamiento medio del solvente, de amplia aplicación en *docking* o durante el análisis de trayectorias de dinámica molecular. También es posible considerar modelos mixtos en los cuales se tienen en cuenta explícitamente determinadas moléculas de agua y el resto se consideran de manera implícita.

1.2.4.3.1.- Modelo de solvente implícito (ISM)

Tanto en los protocolos de *docking* como durante el análisis de trayectorias de dinámica molecular se suelen emplear modelos de solvente implícitos para calcular la desolvatación de receptor y ligando, ya que son lo suficientemente rápidos como para permitir un gran número de cálculos en un corto espacio de tiempo. Sin embargo, es necesario llegar a un compromiso entre exactitud y velocidad, propiedades que suelen estar inversamente relacionadas. Los métodos más populares son el modelo generalizado de *Born* (GB, Onufriev, Case, & Bashford, 2002) o la resolución de la ecuación de *Poisson-Boltzmann* (PB, Fogolari, Brigo, & Molinari, 2002).

El modelo ISM (*implicit solvent model*) deriva de la teoría de los líquidos polares de Deby-Sack (Mehler 1996; Murray and Sen 1996) donde se propone que el efecto de apantallamiento debido al solvente tiene un comportamiento dependiente de la distancia entre las cargas de tipo sigmoidal, siendo la principal contribución a la energía de solvatación aquella que proviene del desplazamiento de la primera capa de moléculas de agua. Los radios atómicos (llamados de Born), cuya parametrización es uno de los principales problemas en este tipo de métodos, miden la distancia efectiva desde un átomo hasta donde empieza el solvente teniendo en cuenta no sólo la disposición del átomo en cuestión sino la de todos los demás. Los radios se calculan mediante una combinación lineal de la parte del átomo expuesta al solvente y la que queda hacia el interior de la proteína. El modelo ISM (Gil-Redondo R., 2006) reduce significativamente la complejidad del cálculo frente a los métodos PB y GB, permitiendo además la descomposición de la energía libre de desolvatación en sus componentes principales: un término carga-carga y las desolvataciones individuales de receptor y ligando.

1.2.4.4.- Interacciones por enlace de hidrógeno

Los enlaces de hidrógeno (HB) son de inmensa importancia para la correcta función de los complejos macromoleculares en los sistemas biológicos. Están involucrados en el mantenimiento del plegamiento de ADN, ARN y proteínas, en el reconocimiento de ligandos y en la estabilidad de los complejos (Chen & Kurgan, 2009). Los HB son interacciones electrostáticas de naturaleza atractiva protagonizadas por un átomo de hidrógeno y dos átomos electronegativos como pueden ser el nitrógeno, el oxígeno o el flúor, que al interaccionar producen el solapamiento en sus nubes electrónicas. El átomo electronegativo unido covalentemente al hidrógeno se denomina donador de enlace de hidrógeno (DHB) y al segundo aceptor de enlace de hidrógeno (AHB) (Figura 12).

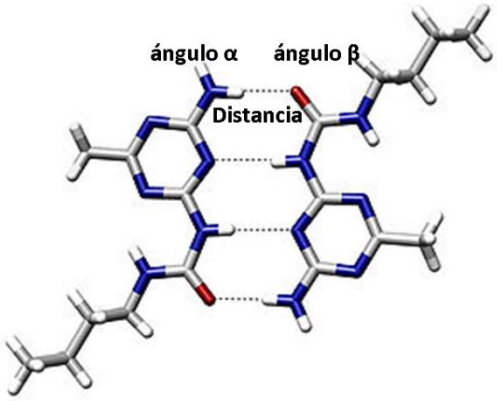


Figura 12. Ejemplo de un enlace de hidrógeno (HB) entre el átomo hidrógeno unido a un átomo de nitrógeno y un átomo de oxígeno definido por 3 variables geométricas: una distancia y dos ángulos [ángulo α (DHB-H…AHB) y ángulo β (H…AHB-X), siendo X un átomo unido a AHB].

Los HB son responsables de la direccionalidad y reconocimiento de substratos y pueden modular la afinidad de la diana por ellos. Su contribución energética depende de los tipos de átomos involucrados en la interacción así como de su geometría preferencial, pudiendo oscilar en rangos de -1 kcal/mol a -40 kcal/mol (Steiner, 2002), de ahí su importancia en el diseño de fármacos (Abraham et al., 2002). A los HB entre grupos de carga opuesta se les denominan puentes salinos.

1.2.4.5.- Interacciones apolares o hidrofóbicas

Las cadenas laterales de ciertos aminoácidos (leucina, valina y prolina), así como regiones de los ligandos, presentan carácter hidrofóbico, esto significa que su interacción con el solvente es muy desfavorable. Si dos centros hidrofóbicos entran en contacto, las moléculas de agua de alrededor son liberadas y esta interacción reduce la energía libre del sistema (efecto

hidrofóbico). El efecto hidrofóbico, entre otras cosas, es una de las fuerzas que estabilizan el plegamiento de las proteínas globulares.

1.2.4.6.- Entropía

El concepto de entropía está íntimamente relacionado con el número de conformaciones microscópicas compatibles con un determinado estado macroscópico. Una molécula aislada es libre de trasladarse, rotar y vibrar. Cuando un complejo intermolecular se forma, algunos de estos movimientos se ven impedidos produciendo una disminución de la entropía. La entropía del soluto (entropía configuracional) se suele dividir en dos partes: conformacional y vibracional. La parte conformacional tiene que ver con la reducción del número de pozos de energía que tanto el ligando como la proteína pueden visitar una vez que ha sucedido la unión, mientras que la parte vibracional se refiere a los movimientos dentro de un pozo de energía en particular. La entropía es difícil de calcular, y se suele ignorar aunque su contribución a la energía libre es importante.

1.2.4.7.- Otras interacciones

Otras interacciones de relevancia, aunque consideradas implícitamente en los tipos anteriormente expuestos, rara vez se tienen en cuenta o si se hace suele ser con aproximaciones muy sencillas. Entre ellas destacamos las interacciones mediadas por moléculas de agua específicas o aguas catalíticas, átomos de halógenos, interacciones entre anillos aromáticos (π-π *stacking,* cuya contribución a la energía libre se estimada va de 0.6 kcal/mol a -7 kcal/mol según el método, (Burley & Petsko, 1985; Jurecka, Sponer, Cerný, & Hobza, 2006)) e interacciones de coordinación de iones metálicos. Estas interacciones pueden ser importantes, determinando en algunos casos la correcta predicción de la unión del ligando.

2.- Técnicas Bioinformáticas

2.1.- Predicción de estructura terciaria basada en relaciones evolutivas: modelado por homología

Conocer la estructura tridimensional de las proteínas nos permite estudiar a nivel molecular cómo funcionan, cómo se comunican o cómo son moduladas. En muchos casos no disponemos de esta información y necesitamos predecirla *in silico* en base a la información de secuencia de la que dispongamos.

Como hemos visto en la introducción, la secuencia de aminoácidos determina en buena medida la estructura tridimensional de las proteínas, al menos de su flexibilidad intrínseca y de las interacciones con pequeño ligandos y otras macromoléculas que pueden modificar su conformación. Por esta razón se puede esperar que sea posible determinar la estructura nativa de la proteína como la estructura representativa del estado de mínima energía libre en condiciones nativas. Sin embargo, a pesar de importantes progresos recientes, estas técnicas basadas en funciones de energía empíricas y llamadas, un poco impropiamente, *ab initio* todavía no garantizan resultados fiables. Afortunadamente, la evolución nos ayuda a extrapolar resultados experimentales y nos proporciona una técnica de predicción muy poderosa llamada modelado por homología.

El modelado por homología se basa en la observación de que, la estructura está determinada por su secuencia de aminoácidos (Anfinsen, 1973), a lo largo de la evolución las estructuras terciarias de las proteínas sufren menores variaciones que sus secuencias de aminoácidos (Chothia & Lesk, 1986), en cuanto secuencias cercanas evolutivamente (*i.e*. proteínas homólogas, descendientes de un ancestro común) se pliegan en estructuras similares y presenten funciones relacionadas. El modelado por homología, o modelado comparativo, es una técnica computacional que explota esta observación para construir modelos atómicos de estructuras tridimensionales de proteínas dianas cuando se conozca la estructura de una o varias proteínas (moldes) relacionada evolutivamente (Cavasotto & Phatak, 2009a).

En un nivel intermedio entre el modelado de estructura basado en funciones de energía que no utilizan moldes y el modelado por homología encontramos las técnicas definidas como *threading* (hilado) en las cuales el molde se identifica evaluando la compatibilidad entre la secuencia diana y un conjunto de posibles moldes con una función de energía empírica (Xu, Jiao, & Yu, 2008). Sin embargo, el aumento de la base de datos tanto de

secuencias como de estructuras y los progresos de los métodos para reconocer secuencias homólogas hacen que en la actualidad las técnicas de *threading* se usen poco.

Las técnicas de predicción de estructura tridimensional, tanto basadas en principios físicos, como las basadas en moldes obtenidos experimentalmente como el modelado por homología, y el *threading*, son herramientas computacionales eficaces para suplir la falta de datos estructurales, llegando en el mejor de los casos a una precisión comparable a la experimental. Se estima que las actuales técnicas de modelado 3D nos permiten construir modelos del 25%-65% de los aminoácidos codificados en los genomas secuenciados, aunque dichos valores difieren significativamente entre genomas individuales (Xiang, 2006a). El 75%-35% restante no son modelables o bien porque no existen aún moldes adecuados para su construcción o bien por su naturaleza desordenada.

La calidad de los modelos obtenidos dependerá de la calidad de los moldes disponibles (por ejemplo la resolución de los cristales de RX), de su distancia evolutiva con la diana de interés medida en términos de identidad de secuencia (ID, *i.e.* el número de residuos iguales entre diana y molde dividido entre la longitud de la región alineada) y de la precisión del alineamiento, posición por posición, de los aminoácidos diana-molde. En términos generales se espera que los modelos generados a partir de secuencias con una identidad mayor del 40% tengan una estructura similar y podrán ser considerados como modelos de alta calidad, mientras que si la identidad es menor del 30% (zona de penumbra o *twilight zone,* Rost, 1999) probablemente presenten estructuras alejadas del molde propuesto y obtengamos modelos de baja calidad.

En los últimos años ha surgido en internet una gran diversidad de servidores automáticos para el modelado por homología, que permiten realizar un proceso de modelado completo sino también descargar estructuras pre-calculadas de bases de datos de modelos por homología generados de modo automático, como por ejemplo SWISS-MODEL (http://swissmodel.expasy.org/SWISS-MODEL.html, Arnold, Bordoli, Kopp, & Schwede, 2006). Con tanta información disponible uno de los problemas a los que nos enfrentamos hoy en día es precisamente la evaluación de la calidad de los modelos para seleccionar los mejores. Es por tanto recomendable el uso de herramientas integradoras como *Protein Model Portal* (PMP, http://proteinmodelportal.org, Arnold et al., 2009) que permitan la comparación de modelos de la misma proteína obtenidos con diferentes métodos.

Otros avances como las mejoras en la capacidad de cálculo y la generación y maduración de métodos y algoritmos (por ejemplo el análisis de perfiles y el mejor

entendimiento y caracterización de la energía que determina la estabilidad de las proteínas) han ayudado a la integración y aplicación de los protocolos de modelado por homología en diversos proyectos de interés biomédicos, destacando entre ellos el estudio de mecanismos de acción (Mueckler & Makepeace, 2008) de mutaciones involucradas en enfermedades (Sun et al., 2007) y el diseño racional de fármacos (Fan, Irwin, & Sali, 2012).

Un protocolo estándar de modelado por homología se puede dividir en las siguientes etapas principales: **(1)** búsqueda de proteínas homólogas (moldes) alineadas contra la secuencia diana, **(2)** selección del mejor molde, **(3)** construcción del modelo 3D, **(4)** refinado y **(5)** evaluación y validación de las estructuras obtenidas (Figura 13).

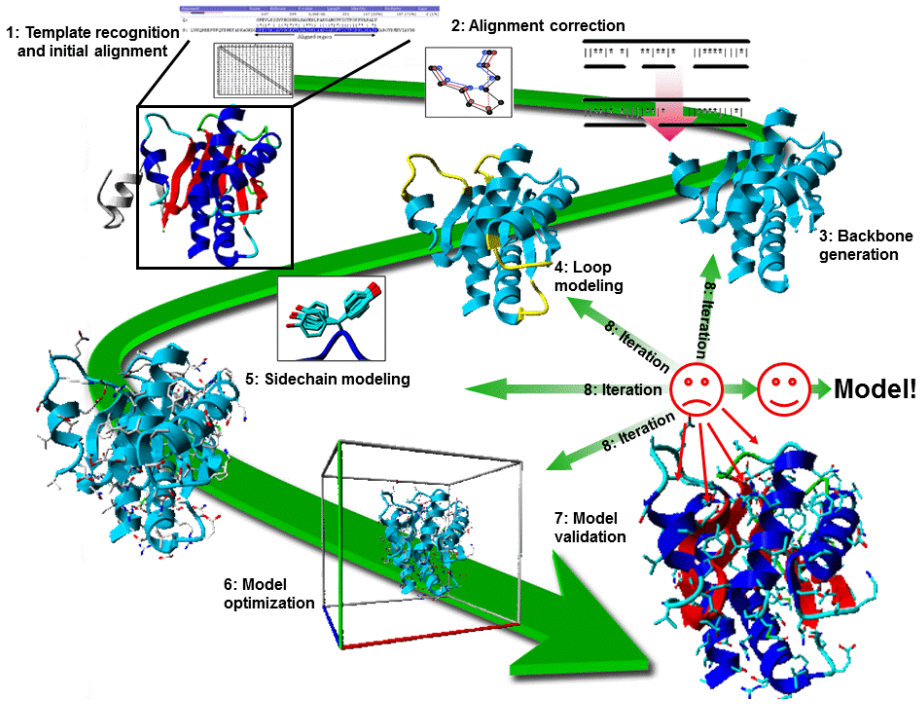


Figura 13. Ciclo típico de modelado comparativo: **1)** reconocimiento del molde y alineamiento inicial, **2)** corrección del alineamiento, **3)** generación del esqueleto proteico, **4)** modelado de bucles, **5)** modelado de cadenas laterales, **6)** optimización del modelo y **7)** validación del modelo (Fuente: Venselaar et al., 2010).

2.1.1.- Alineamiento de secuencias: Búsqueda de homólogos y selección de moldes

En la búsqueda de proteínas homólogas se utilizan algoritmos de alineamiento contra bases de datos de secuencias y/o estructuras. Alinear dos secuencias de proteínas consiste en encontrar la correspondencia entre los residuos de las dos secuencias que tienen más probabilidad de tener un origen evolutivo común, lo cual se hace puntuando su semejanza.

Para entrar más en detalle en los métodos de alineamiento necesitamos definir los conceptos de similitud de secuencia y matrices de sustitución. Tanto la ID como la similitud de secuencia (suma de puntuaciones asociada al tipo de residuos alineados calculadas a partir de matrices de sustitución) se establecen como medidas de evaluación de la calidad de los alineamientos. Las Matrices de Sustitución, de las cuales las más usadas son PAM y BLOSUM (*Percent Accepted Mutations* PAM 1978, *Blocks Substitution Matrix* BLOSUM 1992) resumen la información de las sustituciones entre los diferentes tipos de residuos en grandes bases de datos de proteínas alineadas de manera fiable, y convierten las frecuencias observadas en puntuaciones estadísticas. Las sustituciones entre residuos con propiedades similares (E → D) tienen una mejor puntuación que residuos con propiedades diferentes (E → W) porque se observan con mucha más frecuencia. Como las proteínas a comparar, a pesar de tener un origen común, en general tienen diferentes longitudes debido a eventos de inserciones y delecciones que debemos de tener en cuenta para evitar desfases en el alineamiento, hay que permitir y puntuar la existencia de huecos en el alineamiento o *gaps*. Cabe señalar que en la obtención de la puntuación final o *score* se tienen en cuenta, no sólo las sustituciones entre los residuos sino los eventos de inserciones/delecciones, penalizando la apertura y extensión de los *gaps* incorporados en el alineamiento.

En la Figura 14 se esquematiza un alineamiento con sus principales regiones así como los principales tipos de alineamientos: **(a)** alineamientos por pares, *i.e.* alineamientos entre dos secuencias que tienen como objetivo encontrar la posición relativa entre ambas que maximice el número de coincidencias (medido a través de su ID), solución que ha de satisfacer, al menos en teoría, la correspondencia estructural entre los residuos implicados en el alineamiento, **(b)** los alineamientos múltiples (MSA), *i.e.* la extrapolación del alineamiento por pares al alineamiento simultaneo de un conjunto de secuencias permitiendo la extracción de información evolutiva y la detección de motivos comunes a las secuencias y **(c)** los alineamientos de perfiles, *i.e.* la clasificación y mapeo de las posiciones incluidas en los alineamientos múltiples, tanto de los motivos comunes provenientes de expresiones regulares o segmentos cortos con alto grado de conservación (*match* o coincidencia), como de los elementos sin correspondencia en secuencia (inserciones y delecciones). Dependiendo de la extensión a considerar los alineamientos se pueden clasificar a su vez en locales, construyendo el alineamiento solo para regiones de alta similitud (Smith & Waterman, 1981) y globales, alineando las secuencias completas hasta que se llegue al final de una de ellas (Needleman & Wunsch, 1970), siendo éste el menos efectivo con secuencias altamente divergentes.

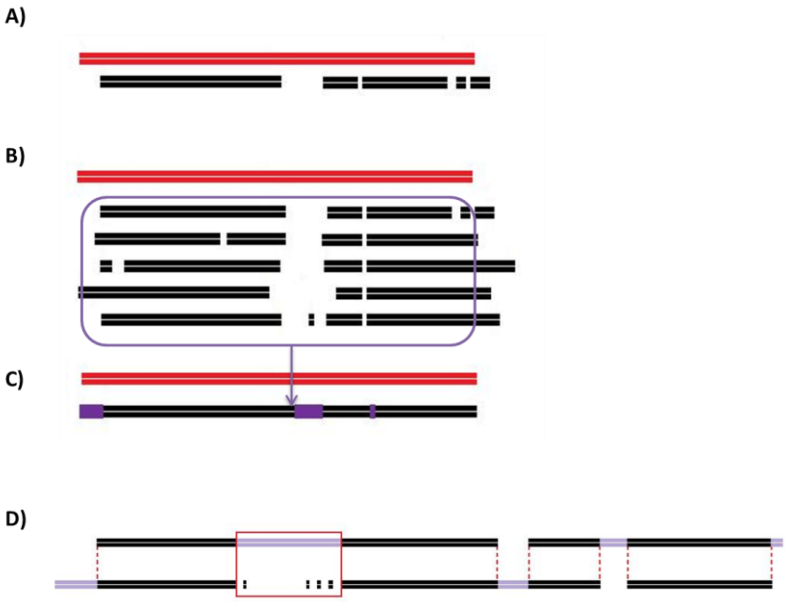


Figura 14. Esquemas de tipos de alineamientos y principales regiones alineadas: **A)** alineamiento por pares de secuencias, **B)** alineamiento de múltiples secuencias (MSA), **C)** alineamiento de perfiles que condensan la información de los MSA y **(D)** tipo de regiones que podemos encontrar en un alineamiento entre pares de secuencias o de perfiles. Zonas de inserciones o delecciones respecto al molde, centrales o en los extremos, y zonas de *gaps* con residuos aislados alineados en su interior. Leyenda: en rojo la secuencia diana, en negro los moldes alineados y en violeta las zonas de *gaps*.

El alineamiento múltiple (MSA) busca alinear *n* secuencia de tal manera que se optimice la puntuación de todos los *n(n-1)/2* pares de secuencias alineadas. Los algoritmos de MSA utilizan una aproximación conocida como alineamiento progresivo, para aproximarse a la solución óptima. Los MSA hacen uso de matrices de sustitución (aunque éstas pueden variar a lo largo del cálculo) e incorporan penalizaciones por apertura y extensión de *gaps*. Por otro lado calculan internamente matrices de distancias, obtenidas a partir de alineamientos por pares independientes, y árboles de distancias que sirven de guía en el alineamiento progresivo.

En los últimos años se ha realizado un gran esfuerzo en el desarrollo de métodos de alineamiento cada vez más y más sensibles como son las búsquedas iterativas (psi-blast (Altschul et al., 1997) y HMMER (http://hmmer.wustl.edu; Eddy, 1998) y los alineamientos perfil-perfil [HHsearch (Söding, 2005) y HHblits (Remmert, Biegert, Hauser, & Söding, 2012)]. Los modelos ocultos de Markov [HMMs, (Krogh, Brown, Mian, Sjölander, & Haussler, 1994)] son un tipo de modelo probabilístico en el que se asume que el sistema a modelar es un fenómeno aleatorio de parámetros desconocidos u ocultos dependiente del tiempo para el cual se cumple una propiedad específica. Partimos de un conjunto finito de estados con sus probabilidades de transición asociadas y un conjunto de variables observables. El objetivo de

los HMMs aplicados al alineamiento de secuencias es el de determinar los parámetros desconocidos a partir de los parámetros observables construyendo un árbol de similitud y comparando los MSA de cada nodo interno del árbol. Son muy efectivos en la detección de patrones conservados entre múltiples secuencias. Por ejemplo ayudan a establecer el consenso de estructura primaria de una familia de proteínas (PFAM). El uso de HMMs permite resumir la información de un MSA en un cadena de símbolos específicos (Match, Insertion, Delection) por posición llamada perfil, cuya precisión se puede incrementar incluyendo homólogos seguros (Sadreyev & Grishin, 2006).

La elección entre un método de alineamiento u otro puede verse condicionado por la ID de las proteínas consideradas. El método más usado como primera opción, porque permite comparar rápidamente una secuencia diana frente a una base de datos con millones de secuencias, es blast (*Basic Local Alignment Search Tool,* Altschul, Gish, Miller, Myers, & Lipman, 1990), pero los resultados de este método son fiables sólo si encuentra homólogos con más de 30% de ID. En el caso de homólogos lejanos, es decir con ID baja, es preferible usar MSA [mafft (Katoh & Toh, 2008), muscle (Edgar, 2004)] o alineamientos de perfiles (psi-blast, hhblits) para garantizar una mejor calidad en el mapeo de las posiciones de los residuos.

2.1.2.- Construcción y refinado de modelos 3D

Una vez tenemos un molde con estructura y el alineamiento molde-diana se construye la estructura tridimensional de la diana proporcionando las coordenadas cartesianas de cada uno de sus átomos. Las estrategias principales en la construcción tridimensional son: **(1)** modelado por satisfacción de restricciones moleculares, **(2)** modelado por ensamblado de fragmentos rígidos, **(3)** modelado mediante coincidencia de segmentos y **(4)** modelado por evolución artificial (Xiang, 2006b).

El método más usado, el modelado por satisfacción de restricciones moleculares, se basa en un procedimiento similar al usado en la determinación de estructuras por RNM. MODELLER (Fiser & Šali, 2003) es uno de los programas más populares que usa este método de construcción y ModBase (http://modbase.compbio.ucsf.edu, Pieper et al., 2011) una base de datos de modelos basados en dicho programa. Al inicio del proceso se establecen una serie de restricciones en la estructura usando el alineamiento como guía: **(1)** restricciones derivadas de la homología como distancias y ángulos diedros, **(2)** restricciones estructurales como preferencias en la longitud y ángulo de enlace, **(3)** preferencias estadísticas por ciertos ángulos diedros y distancias en interacciones no-enlazantes y **(4)** restricciones aplicadas por el usuario tales como la determinación del tipo de estructuras secundarias a construir, etc. Las

restricciones espaciales, expresadas como funciones de densidad de probabilidad, se combinan en una función objetivo que es optimizada usando una combinación de un minimizador por gradientes conjugados y dinámica molecular con *simulated annealing* (ver sección de dinámica molecular). Las restricciones están basadas en la asunción de que las distancias correspondientes entre residuos alineados en molde y diana han de ser similares.

Una vez que tenemos la estructura completa necesitamos optimizar el modelo ya que al realizar modificaciones sobre el molde original, tanto el esqueleto proteico como las cadenas laterales pueden no encontrarse en una posición adecuada. El refinado requiere de una estrategia de muestreo así como una función de energía precisa que guíe en la búsqueda a través del espacio conformacional. El refinado se centra principalmente en el modelado de *loops* (Figura 15) y cadenas laterales.

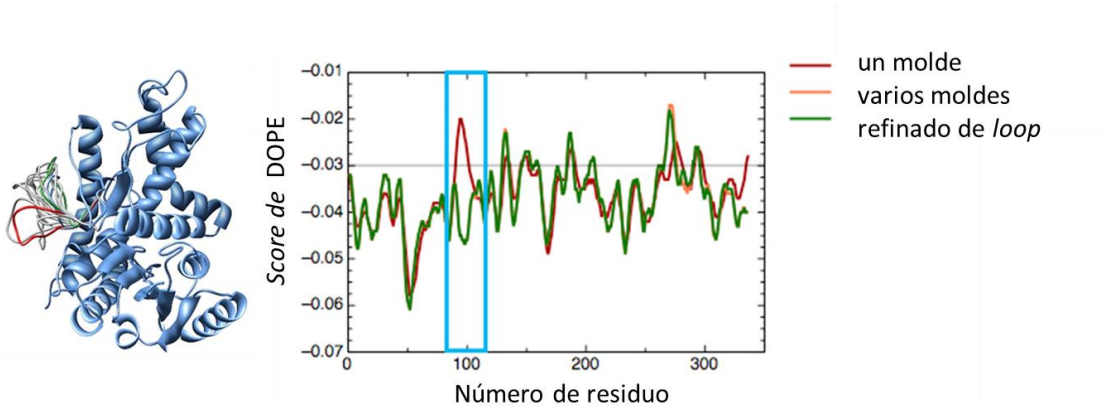


Figura 15. Ejemplo de refinado de estructuras obtenidas con MODELLER: mostramos una proteína diana con una región susceptible de ser mejorada y al lado una comparativa de las puntuaciones DOPE de dicha estructura generada a partir de un molde, de varios moldes y refinado de un bucle o *loop*. La caja azul resalta una región donde el refinado de *loops* mejora la energía asociada a ese fragmento.

Como los errores en el alineamiento son difíciles de corregir en fases avanzadas del refinado se tratan de disminuir mediante un proceso iterativo de re-alineamiento, modelado y evaluación hasta dar con una solución adecuada. Errores en el modelado de proteínas con ID por encima del 40% suele estar relacionados principalmente con la disposición de sus cadenas laterales, sin embargo en el rango de valores de ID entre el 30%-40% los errores en los *loops* adquieren mayor importancia (Sánchez and Sali, 1997). Los *loops* son habitualmente las regiones más variables de las proteínas donde suelen localizarse inserciones y delecciones. Su reconstrucción se realiza usando técnicas *ab initio* y búsquedas en base de datos. Normalmente los *loops* con menos de 12 residuos no presentan un problema de muestreo (Rapp & Friesner, 1999), sin embargo las búsquedas en base de datos se ven limitadas a *loops*

muy pequeños (*i.e.* unos 5-6 residuos) dada su dificultad para encontrar conformaciones cercanas a la nativa en sus librerías de moldes (Fidelis, Stern, Bacon, & Moult, 1994; Deane & Blundell, 2001). Técnicas como la minimización de la energía libre y la dinámica molecular, que serán explicadas en la siguiente sección, pueden ayudar en el refinado de estructuras provenientes de HM sobre todo a nivel de ajustes en cadenas laterales (Fan and Mark, 2004) pero también pueden contribuir a empeorarlo dependiendo de la calidad inicial del modelo generado por HM.

2.1.3.- Evaluación y validación de los modelos 3D obtenidos

Una vez hemos llegado al final del proceso de modelado y refinado, el último paso es evaluar los modelos desde un punto de vista estructural y energético. Es habitual que los modelos por homología contengan algunos errores identificables *a posteriori*. Nuestro objetivo debe ser el de estimar la probabilidad y magnitud de dichos errores así como de establecer, de entre todas las estructuras que podemos construir a partir de los diferentes alineamientos y metodologías, cuál es la más plausible.

2.1.3.1.- Evaluación estructural

Si se conoce la estructura de la proteína diana, es un ejercicio interesante comparar su estructura con los modelos que se han obtenido usando diferentes métodos, con el fin de poder elegir de una manera objetiva el mejor de ellos. Esta evaluación se hace cada dos años en el contexto del *Critical Assessment of Techniques for Protein Structure Prediction* (CASP), en el cual diferentes grupos de investigación intentan predecir la estructura de las mismas proteínas desconocidas usando toda la información disponible en el momento en el PDB. Sin embargo, esta evaluación es en sí una tarea compleja porque los métodos predicen porciones diferentes de la misma proteína, y existen muchas formas de medir la similitud de fragmentos de tamaño variable que pueden llevar a evaluaciones muy diferentes.

El RMSD (ver sección *1.1.2.1.- Variabilidad estructural*) mide la desviación cuadrática media entre los pares de átomos alineados después de la superposición óptima, pero ésta se ve fuertemente influida por las regiones de mucha variabilidad. Por esta razón, en CASP se tiende a usar el test de distancia global (GDT), que hace un promedio ponderado del porcentaje de átomos alineados que se encuentran a una distancia menor de un valor de corte definido (por ejemplo 1 Å, 2 Å, 4 Å y 8 Å) puntuando tanto la extensión del modelo como su precisión.

Si no se conoce la estructura de la proteína diana se pueden evaluar los modelos en términos de la compatibilidad entre sus propiedades estructurales (por ejemplo, longitud y

ángulos de enlace, ángulos de torsión, contactos entre residuos, etc.) y las distribuciones de las mismas propiedades en el PDB.

Nuestra habilidad para predecir correctamente la conformación de las cadenas laterales (rotámeros) viene limitada por la conformación copiada del esqueleto proteico del molde. La obtención de rotámeros colocados de modo incorrecto suele venir causada por residuos mal alineados y/o desplazamientos en el esqueleto proteico. Esto se puede corregir modificando el alineamiento de partida para obtener resultados más precisos o realizar pasos de refinados simultáneos. Cuando encontramos errores en el sitio activo o sus alrededores, debemos reconsiderar el protocolo usado a nivel de la selección del molde y/o su alineamiento. Cuando los errores ocurren lejos del sitio activo, salvo en casos de alostería, nuestro modelo puede seguir siendo válido. Aunque, en general, las medidas de evaluación estructural se consideran parámetros de calidad, son menos útiles a la hora de evaluar y/o comparar modelos que las medidas energéticas, ya que varios de estos factores se han tenido en cuenta durante su construcción.

2.1.3.2.- Evaluación energética

La compatibilidad entre las propiedades estructurales del modelo y de la base de datos de estructuras experimentales se puede traducir en una probabilidad, cuyo logaritmo cambiado de signo se interpreta a veces como una energía libre empírica. Existen muchas funciones de energía libre empírica que se basan en este tipo de procedimiento, considerando propiedades y modelos nulos diferentes para calcular las propiedades estadísticas de las bases de datos. Argumentos tanto de mecánica estadística (distribución de *Boltzmann* de las estructuras posibles de una misma cadena según su energía libre) como evolutivos (distribución de *Boltzmann* de las secuencias con el mismo plegamiento según su estabilidad) sugieren que esta energía empírica está relacionada con la energía libre de la secuencia problema en la estructura examinada (Sippl, 1990). En este marco, la estructura con mejor puntuación energética se identifica con el mejor modelo.

La propiedad más crítica de una función de energía es su capacidad para discriminar los modelos correctos (estructura experimental) de modelos con plegamiento incorrecto (señuelos o *decoys*) lo que tiene un enorme impacto en su capacidad de predicción. Los principales modos de abordar estos cálculos son mediante potenciales estadísticos, mediante cálculos energéticos basados en campos de fuerzas e incluso en los últimos años mediante métodos basados en el aprendizaje automático o *machine learning*. Los potenciales estadísticos, de eficiencia demostrada, son métodos empíricos basados en la frecuencia de

observación de contactos entre residuos, directa o indirectamente (por ejemplo el *score* de DOPE, que será utilizado en los trabajos de investigación 4.1 y 4.2). Los contactos entre átomos o residuos, interacciones energéticamente favorables, son variables que ayudan a discriminar un plegamiento nativo correcto de uno incorrecto. En los modelos se permite un número muy reducido de contactos atómicos infrecuentes ya que las estructuras reales no toleran demasiadas interacciones no favorables. Aunque son computacionalmente económicos, las energías estadísticas no son tan sensibles en la evaluación de estructuras erróneas cercanas a la nativa, especialmente en segmentos como *loops* modelados o cadenas laterales.

2.1.3.3.- Validación de los modelos 3D

Tras la evaluación sólo nos queda la validación experimental. Dado que en todo momento hablamos de simulaciones es necesario probar si nuestra estructura propuesta se apoya de suficientes evidencias experimentales para usarla en los estudios de nuestro interés. Dicha validación suele venir respaldada por datos publicados en literatura como experimentos de mutaciones puntuales de residuos involucrados en la unión de pequeñas moléculas o ligandos a la estructura de la proteína diana , sin embargo en el caso del HM masivo presentado en el artículo 1 no se ha llevado a cabo dicha validación.

2.2.- Predicción de características unidimensionales de las proteínas basadas en la secuencia

Una forma conveniente de representar información estructural es mediante perfiles unidimensionales, que asocian a cada residuo de la proteína un valor numérico (por ejemplo su accesibilidad al solvente) o una etiqueta de estado (por ejemplo etiquetas de estructuras secundarias hélice (H), lámina (E), *loop* (C), o si el residuo se encuentra en un estado desordenado u ordenado). Esta información es útil en la caracterización estructural y funcional de las proteínas. Podemos definir no solo la topología, sino el tipo de ambiente en el que se localiza (regiones globulares o de membrana), identificar modificaciones post-traduccionales como fosforilación de residuos asociados a procesos de regulación, etc. Dada la dificultad en la determinación experimental de ciertos parámetros físicos y la imposibilidad de calcularlos de manera *ab initio* los bioinformáticos han desarrollado algoritmos capaces de predecir estas propiedades usando criterios empíricos que proporcionan una fiabilidad razonable, cercana al 80% en la predicción de estructura secundaria (Leopold & Frank, 2012)

En la presente tesis se ha usado el programa DSSP (Kabsch & Sander, 1983) para la asignación de la estructura secundaria presente en los cristales de RX. Este algoritmo analiza la estructura tridimensional de la proteína, identificando los HB en base a criterios geométricos, y asigna una estructura secundaria a un determinado patrón de HB. Se ha usado además un predictor de estructura secundaria [psipred (Buchan et al., 2010)] basado en alineamientos múltiples generados con psi-blast, un predictor de regiones *coiled-coil* o hélices superenrolladas [ncoil, (Lupas, Van Dyke, & Stock, 1991)], *i.e.* motivos estructurales que consisten en repeticiones de 7 residuos con un patrón determinado que promueven la oligomerización mediante interacciones de 2 a 7 hélices α (Liu et al, 2006), de interés por su papel en transiciones desorden-orden de IDPs, un predictor de hélices transmembrana [TMHMM v2.0, (Krogh, Larsson, von Heijne, & Sonnhammer, 2001)], *i.e.* hélices α insertadas en una membrana que atraviesan a ésta de lado a lado, y un predictor de desorden estructural [disopred2, (Ward, Sodhi, McGuffin, Buxton, & Jones, 2004)] basado en alineamientos múltiples generados con psi-blast y la probabilidad empírica de cada aminoácido de la proteína de estar desordenado sin pertenecer a ninguna clase de estructura secundaria en la forma biológicamente activa de una proteína.. Disopred2 fue entrenado con un conjunto de 715 proteínas provenientes de RX con regiones no resueltas alcanzando una capacidad predictiva del 90% (Jones & Ward, 2003).

2.3.- Dinámica de proteínas y transiciones estructurales

Como hemos visto, las proteínas son moléculas con una estabilidad y flexibilidad característica y variable según la familia a la que pertenezcan y la función que deban realizar. Dichas características se observan a nivel de su dinámica térmica a través de fluctuaciones por residuo, y a nivel de sus movimientos acoplados como cambios conformacionales durante o tras la unión con otras moléculas (Figura 16). Se ha evidenciado que los movimientos acoplados o colectivos son esenciales en las funciones biológicas tales como catálisis enzimática, apertura y cierre de canales, interacciones alostéricas (Nussinov & Tsai, 2013), transducción de señales y reconocimiento dinámico. Los movimientos colectivos pueden ser de diversa magnitud, por ejemplo, movimientos de tipo bisagra, de tipo cizalla, rotaciones de subunidades enteras, movimientos de *loops*, desplegamiento parcial y reajustes en cadenas laterales (Gerstein et al., 1998; Henzler-Wildman et al., 2007). La Figura 16 representa una aproximación temporal de los eventos dinámicos, individuales y colectivos, presentes en proteínas

Existen varias aproximaciones computacionales para caracterizar la flexibilidad de proteínas a partir de su estructura. El método más utilizado para simular el comportamiento dinámico de las macromoléculas es la dinámica molecular (MD, Adcock y McCammon, 2006). Se ha venido empleando de manera satisfactoria desde hace más de 30 años formando parte de estudios de búsqueda de nuevos fármacos, interacciones proteína-proteína y eventos de cooperatividad. Sin embargo, suele resultar más costosa computacionalmente que otros métodos, además de su limitación en el estudio de cambios conformacionales que requieren de periodos muy largos de simulación y no es fácil saber si ha convergido o no. Una de las alternativas más interesantes a la MD para estudiar los movimientos colectivos de las macromoléculas es el análisis de modos normales (NMA). En general, en casos donde se pueden aplicar ambas técnicas parece existir una buena correspondencia entre los movimientos descritos por NMA y MD.

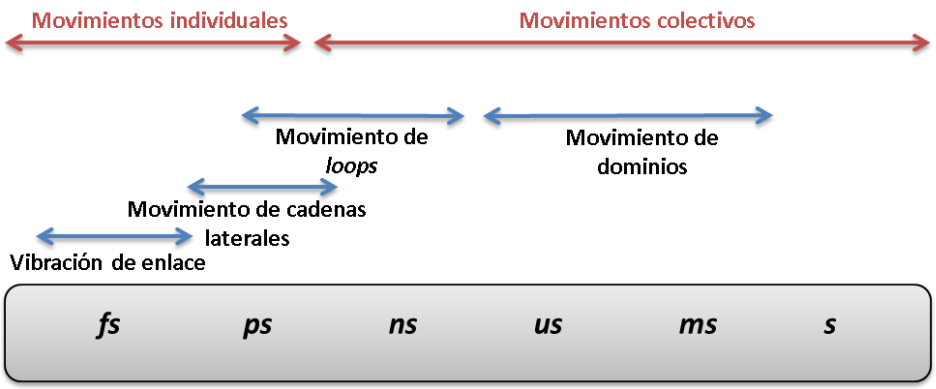


Figura 16. Aproximación temporal de los procesos dinámicos identificados en macromoléculas biológicas.

Como veremos, las técnicas de MD y NMA se pueden aplicar en el ámbito del refinado de estructuras de HM, la predicción de flexibilidad, accesibilidad conformacional y conformaciones alternativas, la predicción de interacciones proteína-proteína y proteína-ligando con relevancia en el diseño de fármacos como inhibidores alostéricos. En la presente tesis nos proponemos estudiar la dinámica y flexibilidad de numerosos complejos usando MD y NMA.

2.3.1.- Dinámica molecular

La dinámica molecular simula la evolución en el tiempo de un sistema de partículas a partir de una configuración inicial (estructuras de RX, RNM o modelado por homología) mediante la integración numérica de las ecuaciones del movimiento de Newton, obteniendo una trayectoria que nos permite analizar el comportamiento del sistema y su estabilidad

alrededor de un mínimo energético. Hace posible la conversión de la información generada a nivel microscópico, como posiciones atómicas y velocidades, a observables macroscópicas, como presión, energía, etc. Emplea potenciales interatómicos detallados y puede considerar explícitamente entornos acuosos, lipídicos, etc., promoviendo que la información obtenida sea lo más realista posible, siempre y cuando el campo de fuerza que se utiliza sea una buena aproximación y la dinámica haya convergido. En la Figura 17se muestra un sistema preparado para MD.

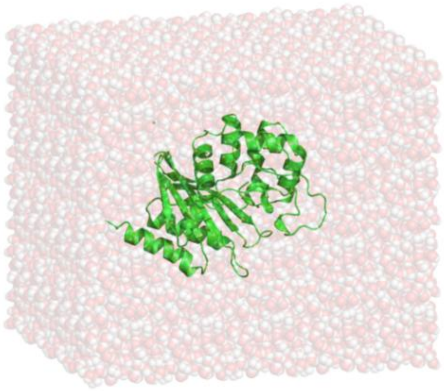


Figura 17. Ejemplo de un sistema para la simulación con MD: proteína globular OXA24 en una caja de aguas usada en la sección de VS [PDB ID: 2JC7].

Del análisis de las trayectorias se puede obtener información estructural y energética muy valiosa para: **(1)** muestreo conformacional, **(2)** descripción de la estabilidad del sistema en equilibrio, **(3)** exploración de su dinámica a lo largo del tiempo y **(4)** refinado de estructuras. A pesar del gran realismo y detalle alcanzados en las simulaciones de MD modelar con esta técnica los grandes cambios conformacionales que tienen lugar en las macromoléculas en el contexto del ajuste flexible no parece ser la opción más eficiente, a no ser que se empleen modificaciones adicionales como la dinámica dirigida (TMD).

2.3.1.1.- Campo de fuerzas

El estado termodinámico de un sistema viene definido por una serie de parámetros como la temperatura (T), la presión (P) y el número de partículas (N), entre otros. Para cada partícula se calculan las fuerzas que actúan sobre ella, así como su posición y velocidad a lo largo del tiempo.

Las simulaciones siguen la aproximación de *Born-Oppenheimer*, que asume que los electrones se adaptan instantáneamente a la configuración del núcleo y por tanto están asociados siempre a su posición, que es la única que se calcula. La fuerza se puede expresar

como $F = m \times a$, donde la m es la masa de la partícula, y a su aceleración. F también se puede expresar como el gradiente de la energía potencial del sistema $F = -\nabla E$, donde $E = \frac{3}{2} NK_B T$ y K_B es la constante de *Boltzmann*. Al conjunto de funciones de potencial y los parámetros necesarios para determinar la energía del sistema a lo largo del tiempo se le denomina campo de fuerzas.

La energía potencial de un sistema es el resultado de la suma de interacciones de pares de átomos enlazados entre sí (energía debido a enlaces covalentes, ángulos y diedros) y las interacciones entre átomos no enlazados (*vdW* y electrostático. Ver sección 1.2.4.- Interacciones no enlazantes). Los campos de fuerza actuales ofrecen un compromiso razonable entre precisión y eficacia en el cálculo numérico. Actualmente existen campos de fuerzas para los tipos generales de moléculas orgánicas: proteínas, lípidos, azúcares, iones, cofactores, etc. Están construidos de forma modular, permitiendo la parametrización de nuevas moléculas o residuos modificados con relativa facilidad en base a los existentes.

Una de las limitaciones de la MD es el tiempo de cálculo. Para acelerarlo se aplican condiciones a dos niveles: **(1)** se establece una distancia de corte a partir de la cual las interacciones entre pares no enlazantes no se calculan, lo cual afecta principalmente al término electrostático y **(2)** se definen condiciones periódicas de contorno (PBC) (Figura 18) permitiendo reducir el número de partículas de solvente del sistema y evitando un efecto abrupto en los bordes de la caja definida para los cálculos.

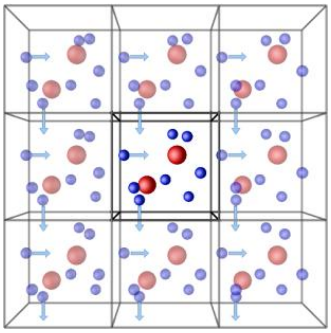


Figura 18. Celda central de simulación y celdas adyacentes representando las condiciones periódicas de contorno.

2.3.1.2.- Etapas y parámetros de un protocolo de simulación por MD

En un protocolo de dinámica molecular se parte de las estructuras obtenidas experimentalmente o computacionalmente y se obtienen sus trayectorias tras pasar por cuatro fases principales: **(1)** minimización de la energía, **(2)** calentamiento y equilibrado, **(3)** producción y **(4)** enfriado. Dependiendo de las propiedades termodinámicas del sistema (N,

energía (E), volumen (V), P y T) que se mantienen constantes a lo largo de la simulación se establecen tres colectivos (*ensembles*) principales: microcanónico (NVE), canónico (NVT) e isotérmico-isobárico (NTP).

A continuación se explican, con un poco más de detalle, cada una de las fases del protocolo.

(1) Minimización de la energía. Aplicando la mecánica molecular, se realiza la búsqueda de conformaciones en mínimos energéticos mediante el ajuste de ángulos, distancia de enlaces y otras variables del campo de fuerzas seleccionado, eliminando posibles fallos en la configuración inicial tales como choques estéricos entre átomos o valores inapropiados en sus ángulos. Los métodos de minimización pueden usar o no gradientes (derivadas) de la energía. Los principales algoritmos son el de máxima pendiente o *steepest descendent* (Cauchy 1847) y el del gradiente conjugado (Hestenes and Stiefel 1952) y suelen usarse de forma conjunta. El método *steepest descendent,* usado como aproximación inicial en sistemas muy alejados de un mínimo energético, utiliza el cómputo de la fuerza para guiar el sistema por el camino más rápido sobre la superficie energética hasta el siguiente mínimo local. El método del gradiente conjugado, al almacenar información de las conformaciones por las que va pasando durante la búsqueda, es más eficiente que el anterior para optimizar la conformación dado un mínimo energético cercano.

(2) Calentamiento y equilibrado. En este paso se establecen las PBC compensando la limitación de espacio definida por la caja de simulación y el efecto de borde (Makov and Payne 1995). Las PBC consideran que los átomos cercanos a la superficie de la caja están sometidos a los potenciales creados por átomos homólogos de las paredes opuestas y viceversa, pudiendo en algún caso cruzar de un lado a otro si esto fuese necesario, reduciendo el efecto de los bordes. Al mismo tiempo, se establece un baño termostático elevando la temperatura del sistema a la temperatura deseada de simulación (a la fase de minimización de la energía le corresponde una temperatura nula), y se observan las propiedades termodinámicas del sistema (P, V y E, por ejemplo). Una vez estabilizadas dichas variables podrá comenzar la fase de producción.

(3) Producción de trayectorias. Una vez que el sistema se encuentra en equilibrio, es decir sus variables permanecen constantes, se pueden obtener instantáneas o *snapshots* de la evolución del sistema a lo largo del tiempo, aportando información sobre la geometría global de la molécula, la localización de sus átomos individuales, así como de su dinámica y estabilidad.

(4) Enfriado. Esta fase reduce la temperatura de la simulación hasta una temperatura seleccionada, obteniéndose así una estructura con alta estabilidad representativa del mínimo energético en el cual se movía la trayectoria. Otra alternativa para conseguir la estructura final de una molécula o complejo consiste en obtener una estructura promedio de un tramo estable de la trayectoria y minimizar su energía para evitar ángulos incorrectos o posibles choques estéricos.

2.3.1.3.- Simulación de membranas biológicas

Las membranas biológicas presentan diversos grados de flexibilidad que dependen de su composición y están relacionados con el desarrollo de su función (Lindahl y Edholm, 2000). Su simulación es un proceso complejo debido a las características estructurales y dinámicas de los fosfolípidos y otros lípidos (por ejemplo el colesterol) que las componen y las proteínas asociadas a ellos (Lindahl & Sansom, 2008).Los fosfolípidos son moléculas de naturaleza anfipática (*i.e.* presentan una parte hidrofóbica y otra hidrofílica). En bicapas lipídicas esta propiedad genera un perfil de presión heterogéneo a lo largo de la membrana (eje Z). La membrana se encuentra en un equilibrio dinámico donde su energía libre varía poco con respecto a su superficie. En este estado, cuando los grupos polares se contraen para evitar la exposición de las colas hidrofóbicas al solvente reduciendo el área de la superficie de la membrana y consecuentemente sus interacciones polares con el solvente, la región hidrofóbica tiende a aumentar su entropía ocupando una área de mayor tamaño (Gullingsrud & Schulten, 2004).Por tanto el área de la superficie de una bicapa lipídica puede variar manteniendo su energía libre y su volumen casi constantes.

Las proteínas de membrana también presentan partes hidrofóbicas e hidrofílicas en diferente proporciones según su nivel de inserción. Por ello, durante las simulaciones por MD de proteínas insertadas en membranas lipídicas es muy importante una adecuada distribución espacial de las regiones polares y apolares. A día de hoy los distintos campos de fuerzas [por ejemplo CHARMM (Lindahl & Sansom, 2008, Jo, Lim, Klauda, & Im, 2009) y AMBER (Case et al., 2005,Skjevik, Madej, Walker, & Teigen, 2012a)] incluyen módulos para el tratamiento de lípidos a nivel atómico, permitiendo refinar estructuras provenientes de HM, analizar interacciones específicas lípido-proteína y estudiar la dinámica de proteínas de membrana en un contexto más realista (Figura 19).

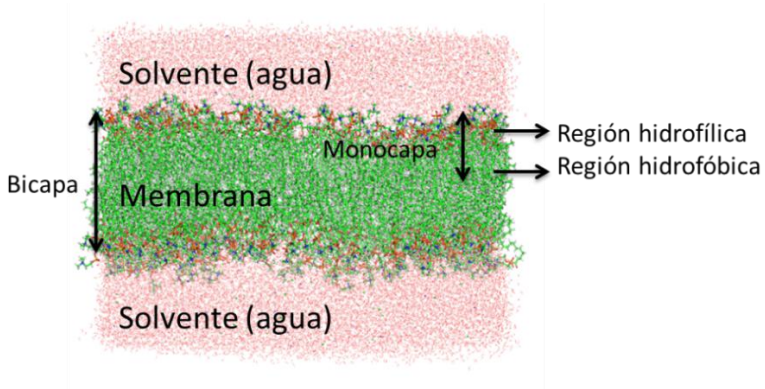


Figura 19. Ejemplo de una membrana preparada para su simulación mediante MD, con sus regiones hidrofóbicas e hidrofílicas claramente diferenciadas.

2.3.2.- Modos Normales

Otra de las técnicas que presentamos para estudiar la flexibilidad de proteínas se basa en la teoría del análisis de los modos normales (NMA) y en el modelo de red elástica (ENM).

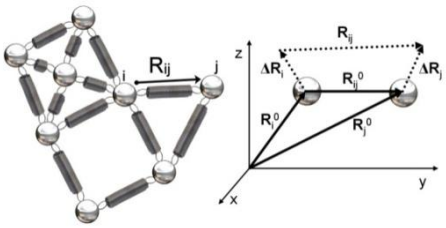


Figura 20. Modelo vibracional de esferas unidas por muelles.

La teoría de los modos normales (NMA), aplicada a las vibraciones moleculares (Wilson et al., 1955) (Figura 20) proporciona una solución analítica completa de las ecuaciones del movimiento de los sistemas sometidos a un potencial de tipo cuadrático (armónico), lo cual representa una buena aproximación si el sistema se mantiene cercano a un mínimo local de la energía. Las soluciones de estas ecuaciones son un conjunto de vectores ortonormales denominados modos normales que representan, en principio, todas las maneras posibles de deformar una estructura alrededor de su conformación de equilibrio, es decir, cualquier movimiento se puede expresar como una combinación lineal de dichos modos. Desde el punto de vista de la mecánica estadística esta normalidad es importante porque implica que los modos normales son independientes, y simplifica los cálculos analíticos de las cantidades termodinámicas. Los modos normales, a diferencia de la MD, proporcionan una solución analítica simple de la dinámica térmica de las proteínas en su estado nativo, lo cual quiere decir que no nos tenemos que preocupar de la convergencia de los cálculos termodinámicos, y

suelen presentarse como un método computacional efectivo para el estudio de los movimientos colectivos de gran amplitud (Bahar, Lezon, Yang, & Eyal, 2010).

En los años 90 el uso de los NMA ganó popularidad en la descripción de la dinámica de proteínas gracias a la aparición de los modelos de red elástica [ENM, (Tirion, 1996)], que predicen las fluctuaciones térmicas de una proteína usando sólo su estructura nativa. El campo de fuerza de los ENM se construye en base al principio de mínima frustración, que supone que todas las interacciones existentes en la estructura nativa de la proteína son ventajosas energéticamente. Además, para reducir al máximo el número de parámetros, se asigna a todas las interacciones nativas la misma constante de fuerza que describe cómo aumenta la energía cuando los átomos que interaccionan se alejan de la conformación más favorecida. De esta manera, los ENM sólo varían en base al criterio que se usa para decidir si un par de átomos *ij* interaccionan ($C_{ij} = 1$) o no ($C_{ij} = 0$), y su campo de fuerza se define como $\mathrm{V} = \frac{k}{2}\sum_{i<j} C_{ij}\left(r_{ij} - \bar{r}_{ij}\right)^2$ donde r_{ij} es la distancia entre los átomos *i* y *j*, $\bar{r}_{ij}$ es su posición de equilibrio, determinada por la estructura experimental, *k* es la constante de fuerza y C_{ij} determina si los átomos están interaccionando o no. Este modelo tiene la ventaja de que sus resultados son robustos, no dependiendo de los parámetros de un campo de fuerza como en MD y de los tiempos de cálculo, y se elimina la costosa necesidad de conducir la estructura atómica hacia un mínimo energético, lo cual a veces puede alejarla de la estructura experimental. Sin embargo, el campo de fuerza del ENM carece de detalles moleculares, como la descripción de las cargas eléctricas en la molécula, las interacciones con el solvente y la membrana etc., y además se basa en la aproximación armónica que es válida sólo muy cerca del estado de equilibrio. Por estas razones, sus resultados son fiables sólo para los modos colectivos de baja frecuencia y gran amplitud que son más robustos respecto a los detalles moleculares porque a lo largo de estos modos la energía varía poco incluso por desplazamientos grandes.

Las características que definen los diferentes ENM que han sido propuestos son el criterio usado para decidir si dos átomos interaccionan, que puede estar basado sólo en los C_α, carbonos β (C_β, el primer carbono de la cadena lateral unido al C_α) o, más raramente, en todos los átomos pesados, y especialmente los grados de libertad que se usan, entre los cuales se encuentran los siguientes: **(1)** las distancias de los *N* átomos representativos (típicamente C_α) de su posición de equilibrio Δr_i que, junto con la hipótesis de que las fluctuaciones son isótropas, constituyen el modelo de red gaussiano (GNM, Bahar et al., 1997), el ENM más simple con N grados de libertad; **(2)** las coordenadas cartesianas de los átomos representativos

(3N grados de libertad) que definen el modelo de red anisotrópico (ANM, Atilgan et al, 2001) y **(3)** los ángulos de torsión de la cadena principal (2 grados de libertad por residuo), que definen el modelo de red torsional (TNM, Méndez and Bastolla, 2010). El hecho de usar grados de libertad torsionales permite fijar las longitudes y los ángulos de los enlaces covalentes que son mucho más rígidos, reducir el número de grado de libertad respecto al ANM y al mismo tiempo representar en manera bastante precisa la dinámica de los átomos del *backbone*, N-C_α-C_β-C-O en vez que sólo los C_α. De esta manera, en el TNM se usan todos los átomos para calcular las interacciones intramoleculares y la energía cinética. En este trabajo se ha usado el modelo TNM para cálculos de modos normales.

En la aproximación armónica, las energías cinética (*K*) y potencial (*H*) están representadas con una forma cuadrática de los desplazamientos, y la diagonalización simultánea de las matrices que las definen permite obtener los modos normales de vibración de la macromolécula, es decir, los autovalores (frecuencias) y los autovectores (amplitudes) propios del sistema. Los modos normales de baja frecuencia obtenidos de esta manera pueden describir movimientos colectivos de la proteína de tipo bisagra, giro, torsión, etc. (Figura 21).

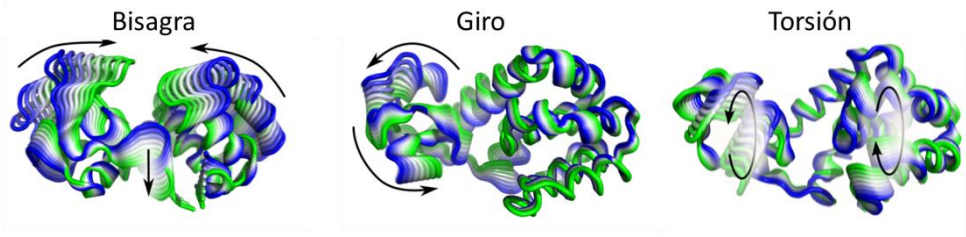


Figura 21. Ejemplos de movimientos descritos por los modos normales de una proteína.

Según el teorema de equipartición, la energía en un sistema armónico en equilibrio térmico se reparte por igual entre todos los grados de libertad. Como la energía media asociada al modo normal α es $\frac{1}{2}\boldsymbol{\omega_\alpha^2 A_\alpha^2 = kT/2}$, donde ω_α^2 es la frecuencia y A_α^2 es la amplitud, esto implica que las frecuencias más bajas tienen amplitud mayor y contribuyen más a los desplazamientos. Los movimientos descritos por estos modos normales de baja frecuencia son movimientos colectivos que involucran a un gran número de residuos, como movimientos de dominios. Se ha demostrado que la mayoría de los cambios conformacionales que experimentan las proteínas se pueden representar bien usando sólo unos pocos modos de baja frecuencia (Hinsen et al., 1999; Marques y Sanejouand, 1995). Sin embargo, a pesar del gran número de estudios de NMA llevados a cabo, no es fácil identificar qué modos son funcionalmente relevantes sin datos experimentales adicionales. Por ello, Hub y Groot proponen un nuevo término denominado análisis de modos funcionales (FMA, Hub y Groot,

2009) cuyo objetivo es explicar las variaciones de lo que denominan *cantidad funcional*, un valor que puede incluir características como la apertura de un canal, la geometría del sitio activo o la cavidad accesible al solvente, en términos de movimientos colectivos que maximicen su relación con esa variable.

2.3.2.1.- Validación del análisis de modos normales

La aplicación de NMA al estudio de la dinámica de proteínas puede parecer arriesgado *a priori*, en particular debido a las limitaciones asociadas al uso de una aproximación armónica mientras analizamos movimientos de gran amplitud, y debido a la falta de precisión molecular y a la ausencia de solvente en el campo de fuerza ENM. Sin embargo, estudios comparativos de NMA con datos de estructuras experimentales y simulaciones de MD validan el uso de NMA incluso con modelos ENM (Rueda et al, 2007; Skjaervena et al, 2009). En estas validaciones se suelen comparar las fluctuaciones predichas por los ENM con **(1)** las fluctuaciones atómicas de estructuras cristalográficas, *i.e. B-factors*, **(2)** las fluctuaciones que se observan en trayectorias de MD o **(3)** los conjuntos de conformaciones obtenidas por RNM.

La correlación de las predicciones de NMA con los *B-factors* suele ser bastante buena (0.55-0.59) (Brooks y Karplus, 1983; Eyal, et al.,2007; Go, et al., 1983; Yang, et al., 2006), a pesar de que los *B-factors* contienen información sobre los grados de libertad de cuerpo rígido de las proteínas, que no se pueden tener en cuenta en la NMA, y además suelen estar condicionados por restricciones de empaquetamiento y desorden cristalino (Halle, 2002; Hinsen, 2008; Soheilifard, et al., 2008) o por la incorporación de errores durante el refinamiento (Carugo y Argos, 1999). Las técnicas de dinámica esencial (ED, Yang, et al., 2007), también llamadas análisis quasi-armónico, consisten en obtener coordenadas colectivas a través del análisis de componentes principales (PCA, una transformación lineal que reduce la dimensionalidad de los datos y nos permite ordenarlos en base a su variabilidad) de las trayectorias de MD. Las coordenadas de ED revelan un gran parecido con los espacios conformacionales descritos con NMA (Rueda et al, 2007). Por último, las técnicas de RNM proporcionan un conjunto experimental de conformaciones representativas del *ensemble* de equilibrio de la proteína. El análisis PCA de estas conformaciones permite verificar experimentalmente los movimientos colectivos predichos por ENM (Bahar and Rader, 2005).

2.4.- Interacciones moleculares: *docking* y cribado virtual

Cuando las estructuras de la proteína diana y el ligando son conocidas, el anclaje o *docking* molecular es la técnica más usada para predecir las interacciones que se establecen entre ellos. La extrapolación del *docking* a quimiotecas con un gran número de moléculas pequeñas (desde unos cientos hasta millones) se conoce como cribado virtual de quimiotecas o *virtual screening* (VS,Figura 22)

El VS de ligandos representa una alternativa computacional eficaz para reducir costes limitando el número de candidatos potencialmente activos que han de ser ensayados experimentalmente (Shoichet, 2004). El objetivo del VS es seleccionar computacionalmente aquellos compuestos (ligandos) que tienen mayor probabilidad de interaccionar satisfactoriamente con una diana biológica, generalmente una enzima o un receptor, utilizando sus estructuras atómicas. Debe ser capaz, además, de diferenciar entre moléculas deseadas (activos) y no deseadas (señuelos o *decoys*) dentro de la quimioteca. La técnica masiva del VS va ganando relevancia y se está convirtiendo en una fuente de moléculas *lead* en la búsqueda de nuevos fármacos. El VS presentado en esta tesis se ha basado en la estructura experimental de una proteína y se ha llevado a cabo en la plataforma de cribado virtual VSDMIP (Cabrera et al, 2011), que facilita enormemente el almacenamiento y manejo de los resultados.

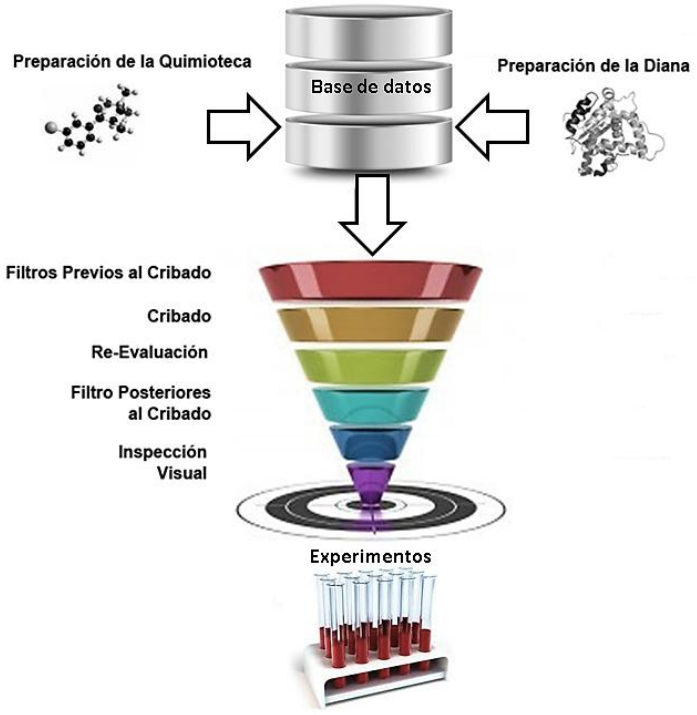


Figura 22. Protocolo estándar de *docking* y cribado virtual que incluye desde la preparación de las moléculas hasta la selección de los mejores candidatos o *hits*.

Una de las medidas que se utilizan a la hora de seleccionar los mejores candidatos es la eficacia del ligando (LE), *i.e.* la energía libre de unión del complejo proteína-ligando dividida entre el número de átomos pesados del ligando. Para evaluar la capacidad predictiva (Ecuación 7) de un programa de docking u otro programa bioinformático, la medida más adoptada es el área bajo la curva ROC (AUC, Hanley et al, 1982). La curva ROC (Figura 23) es una representación gráfica de la tasa de verdaderos positivos o TP, *i.e.* la sensibilidad, frente a la tasa de falsos positivos, *i.e*. la especificidad, para un sistema clasificador binario (por ejemplo la clasificación de las soluciones de docking en activos e inactivos), según se varía el umbral de discriminación. Cuando el AUC es igual a 0.5 el clasificador se comporta como un clasificador aleatorio. Cuanto mayor sea el AUC, el programa de docking en este caso, discrimina mejor los TP de los FP, donde 1 significaría que es un clasificador perfecto con sensibilidad para todos los umbrales (Figura 23).

Ecuación 7: $\text{Precisión} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$

Ecuación 8: $\text{Sensibilidad} = \frac{\text{TP}}{\text{TP+FN}}$

Ecuación 9: $\text{Especificidad} = \frac{\text{TN}}{\text{TN+FP}}$

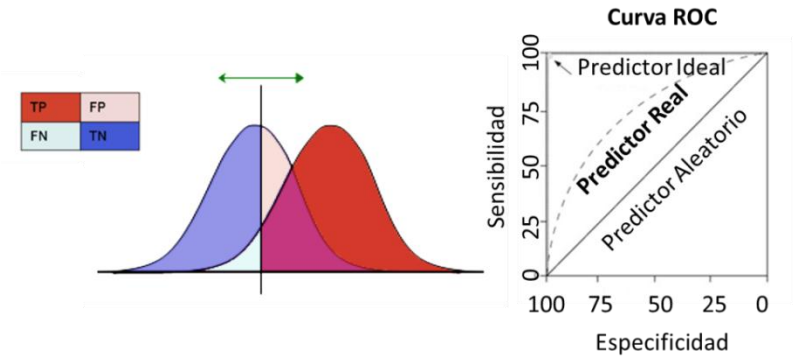


Figura 23. Evaluación de los resultados de un clasificador/predictor**. A)** Clasificación: verdaderos positivos (TP), falsos positivos (FP), falsos negativos (FN), verdaderos negativos (TN), **B)** área bajo la curva ROC (AUC) y capacidad predictiva.

En lo que se refiere a *docking*, la parte más importante de la estructura de la proteína es la zona de unión del ligando, conocido como bolsillo o cavidad de unión, que en el caso de enzimas coincide con el sitio activo de la proteína. La unión del ligando en dicha cavidad puede activar o inhibir la respuesta fisiológica de la diana. Su localización puede ser superficial o estar enterrada en el interior de la proteína y normalmente se forma al mismo tiempo que ésta se pliega. La disposición 3D de las cadenas laterales y del esqueleto de los residuos de la proteína determina la especificidad del ligando por esa diana en particular. Por último, algunos residuos

del centro activo se han conservado a lo largo de la evolución, principalmente aquellos relacionados con la actividad de la diana. Conociendo dónde está situado el sitio activo podemos guiar al algoritmo de *docking* a una región en particular de la proteína en vez de buscar en toda la proteína (*docking* ciego).

Hay dos elementos clave en un protocolo de *docking*: **(1)** una componente estructural o función de muestreo que explore todas las posibles poses del ligando en el sitio activo del receptor y **(2)** una componente energética o función de puntuación o *scoring* que priorice las poses en concordancia con su ajuste preferente al sitio activo.

Para tratar millones de moléculas con este método es necesario hacer ciertas aproximaciones y encontrar el balance adecuado entre velocidad de cálculo y precisión de los resultados. Una de ellas es la discretización del espacio conformacional, al considerar proteína y/o ligandos como entes rígidos durante la fase de muestreo con un número reducido de conformaciones representativas. Otras aproximaciones surgen a la hora de identificar y evaluar correctamente las interacciones que ocurren cuando ligando y receptor forman el complejo, durante la evaluación y *ranking* con la función de *scoring* elegida. De entre todas ellas, los enlaces de hidrógeno son uno de los factores más importantes en la formación y estabilidad de los complejos moleculares y permiten una correcta disposición de las moléculas para la realización de su función biológica, siendo, por ello, clave en el proceso de diseño de fármacos asistido por ordenador.

Las moléculas orgánicas cuentan con un elevado número de grados de libertad que dan lugar a la explosión combinatoria del número de posibles conformaciones. Existen diversos métodos que persiguen la reducción del espacio conformacional con el fin de buscar las poses más representativas de la unión de los complejos. Según el nivel de flexibilidad que permitamos durante un protocolo de *docking* así se establecen tres aproximaciones diferentes: *docking* rígido, *docking* con proteína rígida y ligando flexible y *docking* flexible. El seleccionar una u otra dependerá del proyecto y del tiempo disponible. Cuanta mayor precisión necesitemos para definir la flexibilidad de las moléculas, más caro será computacionalmente. Sin embargo, hay que destacar que los métodos actuales no permiten resultados precisos para el *docking* completamente flexible, lo que constituye una de las fronteras a explorar por estas técnicas.

2.4.1.- Evaluación estructural

Para evaluar los métodos de *docking* se suelen usar complejos proteína-ligando determinados experimentalmente, separándolos en proteína y ligando y volviéndolos a juntar mediante

docking rígido o flexible con el algoritmo en cuestión. Después, se calcula el *RMSD* entre pares equivalentes de átomos en cada pose del *docking* y en la estructura experimental. Valores de *RMSD* por debajo de 1.0 Å representan resultados aceptables, aunque muchos autores aumentan este límite o *cut-off* a 1.5 Å ó 2.0 Å. Por encima de este valor los resultados se consideran incorrectos.

Una medida que muestra la eficacia del método de *docking* es el porcentaje de acierto o éxito, definido como el porcentaje de estructuras predichas con valores de *RMSD* ≤ 2.0 Å. El valor medio de este parámetro para la mayoría de los programas de *docking* en general no supera el 75 % (Kellenberger, Rodrigo, Muller, & Rognan, 2004).

2.4.2.- Evaluación energética

Una función de *scoring* o de puntuación es una ecuación matemática que cuantifica la fuerza con la que un ligando se une a su proteína. Ésta debe ser capaz de colocar la estructura experimental cerca del mínimo global, o por lo menos de un mínimo local del paisaje energético del complejo representado como un espacio multi-dimensional con un gran número de valles y montañas. Si esto se cumple, la función de *scoring* sería capaz de distinguir y elegir como mejor solución aquella pose con el menor valor de *RMSD* con respecto a la estructura experimental. Generalmente se asume que el problema de muestreo (generar mediante el algoritmo de docking una pose correcta según el criterio de *RMSD* indicado arriba) está más o menos resuelto, pero la baja correlación encontrada entre los valores de energía calculados con la función de *scoring* y el RMSD plantea una limitación en la capacidad de reconocer la pose correcta y promocionarla respecto al resto de poses evaluadas.

Asimismo, al emplearse en cálculos que requieren de un gasto computacional elevado, la función de energía debe presentar un equilibrio razonable entre precisión y rapidez. Los tres tipos principales de funciones de evaluación energética utilizadas durante el *docking* (de puntuación o *scoring* y de orden o *ranking*) se diferencian en los datos usados en su derivación: empíricos, basados en el conocimiento y basados en campo de fuerzas.

(a) Empíricas

Se basan en la regresión multi-lineal entre medidas experimentales de actividad y características relevantes en la formación de los complejos como son los enlaces de hidrógeno, las interacciones iónicas y los contactos polares y no polares. La contribución de cada una de estas interacciones a la energía de unión global viene ponderada por un coeficiente obtenido de un conjunto de prueba (*training set*) y por tanto su aplicación es limitada fuera de dicho conjunto.

(b) Basadas en el conocimiento

Partiendo de la información de estructuras 3D experimentales se establecen los tipos de interacción que suceden habitualmente entre ligandos y dianas y su frecuencia, relacionando éstas con su energía libre de una manera parecida a como se derivan los campos de fuerza para el plegamiento a partir de grandes bases de datos de estructuras. Una de las limitaciones de las funciones basadas en el conocimiento está relacionada con la definición de un estado de referencia en el que no exista la interacción bajo estudio, lo que es difícil de establecer.

(c) Basadas en campos de fuerzas

Las funciones de *scoring* basadas en un campo de fuerzas descomponen la energía de interacción del ligando en la suma de las interacciones atómicas individuales, siendo éstos los términos de energía de *vdW*, energía electrostática, energía de enlace torsional/rotacional/vibracional, etc., que se emplean en los campos de fuerza de mecánica molecular. Los mismos problemas presentes en el campo de fuerzas se presentan en la función de *scoring* en la medida en que sus resultados dependen de lo buenos que sean los parámetros.

2.4.3.- Diseño de fármacos: plataforma de cribado virtual VSDMIP

La unidad de bioinformática del CBMSO (http://ub.cbm.uam.es/), en su línea de investigación de diseño de fármacos asistido por ordenador dirigida por el Dr. Antonio Morreale, ha desarrollado y puesto a disposición de la comunidad científica diferentes herramientas para el estudio y predicción de interacciones moleculares englobadas en la plataforma modular de cribado virtual VSDMIP (Cabrera et al., 2011).

VSDMIP incorpora: **(1)** la preparación de ligandos que consiste en la generación de confórmeros a partir de cadenas de SMILES (Weininger, 1988), *i.e.* una notación lineal para codificar estructuras moleculares, y el cálculo de sus cargas usando MOPAC (Stewart, 1990); **(2)** la preparación de la proteína que engloba: **2a)** añadir átomos de hidrógeno y otros átomos que falten, **2b)** pre-cálculo de los valores de la energía de interacción (*vdW* y coulómbico) entre un determinado tipo de átomo de la proteína y del ligando en puntos discretos del espacio 3D, generando mallas o *grids* de energía de interacción con el fin de acelerar los cálculos. Estas *grids* se generan con el programa CGRID (Pérez and Ortiz 2001) en una caja tridimensional definida en una zona de la superficie de la proteína, normalmente un sitio activo conocido. Su función de *scoring* (Pérez & Ortiz, 2001)) está basada en el campo de fuerzas de AMBER y contiene los términos de *vdW* y electrostático, **(3)** el anclaje molecular

mediante el programa CDOCK (Pérez and Ortiz 2001), un programa de *docking* basado en DOCK (Kuntz et al. 1982) que realiza búsquedas exhaustivas manteniendo rígidos proteína y ligando y variando sus grados de libertad relativos (traslaciones y rotaciones), haciendo uso de las *grids* pre-calculadas de energía potencial para obtener su energía de unión. La flexibilidad del ligando se incorpora ejecutando estos cálculos sobre un conjunto de confórmeros del ligando pre-generados con el programa ALFA (Gil Redondo R., 2006). Una de las principales ventajas que presenta este programa de *docking* respecto a otros es su rapidez. El pre-cálculo de *grids* de las proteínas y, sobre todo, de los confórmeros del ligando permite reducir los tiempos de cálculo respecto a programas de *docking* de uso tan extendido como AutoDock (Morris et al. 1998) manteniendo una capacidad predictiva y precisión comparables, lo que supone una prometedora aplicabilidad en protocolos de VS, donde el tiempo de cálculo por molécula es un factor limitante; **(4)** el refinado de poses. Un algoritmo tipo SIMPLEX se encarga de minimizar la posición de las poses mediante movimientos de cuerpo rígido, es decir, se permiten los movimientos globales de traslación y rotación mientras los grados de libertad internos del ligando permanecen fijos y **(5)** el r*anking*. Después de la aplicación del SIMPLEX la puntuación final se re-asigna usando una función de *scoring* más completa que incorpora tanto las desolvataciones de receptor y ligando como una corrección energética para los enlaces de hidrógeno (Morreale, Gil-Redondo, & Ortiz, 2007).

2.4.4.- Optimización *hit-to-lead*

Cuando diseñamos nuevos fármacos, los ligandos seleccionados deben poseer ciertas características importantes como la especificidad hacia una determinada diana terapéutica además de una correcta selectividad para que no se una a otras proteínas con cavidades similares, lo que produciría efectos secundarios. Por ello, una vez tenemos los candidatos o *hits* con eficacia probada experimentalmente, se pasa a un proceso de optimización de dichas moléculas con el fin de potenciar su actividad y especificidad mejorando sus interacciones y reduciendo la cantidad de fármaco que se deberá administrar a los futuros pacientes. Otro de los objetivos de la optimización de los compuestos se centra en la mejora de alguna de sus propiedades químicas como es la solubilidad.

2.5.- Terapia personalizada: *microarrays* de expresión

La variabilidad genética entre individuos de una misma especie y entre diferentes especies es extensamente conocida. Las mutaciones puntuales, inserciones y delecciones producidas a nivel de genoma pueden alterar en mayor o menor medida la estructura y la dinámica de las

proteínas que codifican dependiendo de su localización espacial y su importancia durante el ejercicio de su función. Diversas técnicas experimentales, como son los *microarrays* de expresión de oligonucleótidos que veremos en esta tesis, pueden ayudar en la caracterización de la variabilidad genética de los individuos con el fin de diseñar terapias personalizadas.

La principal aportación de la técnica de los *microarrays* fue su capacidad de obtener una gran cantidad de medidas simultáneas de una muestra en las mismas condiciones experimentales. Dicha técnica provocó un cambio de paradigma en la biología molecular de la era post-genómica, del estudio detallado de un determinado gen al estudio global aunque con menor resolución de numerosos genes simultáneamente. Diferentes tipos de *microarrays* (de ADN, ARN, proteínas, de expresión, etc.) se han venido aplicando en los últimos años asociados a: **(1)** numerosos estudios de genes con expresión diferencial entre varias condiciones (sanos/enfermos, mutantes/salvajes, tratados/no tratados), **(2)** la clasificación molecular en enfermedades complejas, **(3)** la identificación de genes característicos de una patología, **(4)** la predicción de respuesta a un tratamiento y **(5)** la detección de mutaciones y SNPs, entre otros (Trevino, Falciani, & Barrera-Saldana, 2007).

La técnica de *microarrays* de ADN combina la sensibilidad de la reacción en cadena de la polimerasa (*polimerase chain reaction* o PCR) con la selectividad de la hibridación de las hebras de la doble hélice del ADN uniendo de modo complementario los pares adenina- timina (AT) y citosina-guanina (CG) mediante la formación de HB característicos entre ellos (2 ó 3 HB respectivamente). Las muestras a analizar se hibridan con sondas de secuencia conocida previamente unidas al soporte del *microarray* produciendo fluorescencia medible experimentalmente. Sin embargo, dado que la contaminación de una muestra también suele emitir fluorescencia o que las sondas pueden hibridar parcialmente, se requiere de un filtrado y normalización de los datos *a posteriori* para su correcta interpretación. Esta técnica presenta grandes ventajas respecto al clonaje molecular a nivel de costes económicos y de tiempo, pero también algunos inconvenientes como: **(1)** una menor capacidad de identificar correctamente una secuencia (FP y FN) debido a hibridaciones parciales o a características específicas de cada sonda (temperatura de hibridación diferentes, etc.) y **(2)** la complejidad en el diseño de las sondas y su corta vida en el caso de organismos con altas tasas de mutación.

2.5.1- Análisis de datos provenientes de *microarrays*

Se han diseñado numerosos programas para el análisis de *microarrays* de DNA, sin embargo dichos programas no son transferibles al análisis de datos de otras plataformas/protocolos diferentes o requieren de grandes ajustes para su utilización. Uno de los mayores cuellos de

botella para la extensión de estas técnicas es precisamente el análisis automático y generalizado de los datos obtenidos.

Varios factores, tales como las diferentes tecnologías y plataformas, los criterios estadísticos adoptados, los protocolos y variabilidad de laboratorios, influyen en la concordancia entre los diferentes estudios publicados, haciendo difícil su comparación. El pre-procesado de los datos mediante su normalización, tanto intra-*microarray* como inter-*microarray* juega un papel importante en la reducción del ruido producido por dichos factores.

El análisis de datos de *microarrays* puede englobar el análisis de imagen, la visualización y expresión diferencial, el análisis de componentes principales o PCA, el *clustering* o agrupamiento de datos, la clasificación y el análisis de regresión y de supervivencia (Ying & Sarwal, 2009). Las características particulares de los conjuntos de datos a analizar, así como las preguntas a responder, hacen que seleccionemos un tipo de algoritmo y filtros diferentes según el caso.

Algunos de los métodos más usados están basados en el *clustering* de soluciones de modo jerárquico, *i.e.* agrupa los resultados asociados a un árbol filogenético (por ejemplo *average-linkage clustering*), o no jerárquico, *i.e.* agrupa los resultados en base al conocimiento *a priori* del número de *clusters* o grupos representados en los datos (por ejemplo *k-means,* obteniendo el número de *clusters* por otros métodos como PCA). La clasificación basada en el *clustering* es muy dependiente del algoritmo usado, el modo de normalizar los datos y de la medida de similitud usada durante el proceso. Una alternativa a estos métodos cuando disponemos de información previa acerca de qué genes deben estar agrupados es la de los métodos supervisados como SVM o *support vector machine* (Quackenbush, 2001).

Las discrepancias en los resultados de los *microarrays* son consecuencia de las diferencias en las medidas obtenidas con un clasificador como son la precisión, *i.e.* el grado de conformidad de la cantidad medida respecto a su valor real (Ecuación 7 en apartado anterior), la sensibilidad, *i.e.* el rango de concentración de moléculas de la diana en el cual podemos realizar medidas precisas , la reproducibilidad, *i.e.* el grado al cual medidas repetidas de la misma calidad muestran resultados iguales o similares, y la especificidad, *i.e*. la habilidad de una sonda para proporcionar una señal influenciada exclusivamente por la molécula diana (Ecuación 8 en apartado anterior).

En nuestro caso, más que buscar patrones de expresión diferencial como es habitual, pretendemos realizar la identificación de la presencia/ausencia de una mutación o inserción así como la estimación de su nivel de detección.

3.- Objetivos

La presente tesis está enfocada a mejorar la caracterización de las interacciones proteína-fármaco y la capacidad predictiva del *docking* y el VS para su aplicación en protocolos de diseños de fármacos asistidos por ordenador basados en estructura. Profundizaremos en :**(a)** el modelado de las estructuras 3D de las proteínas y en la predicción de desorden por su factor limitante en la aplicabilidad del *docking*, **(b)** la caracterización de la dinámica intrínseca de las proteínas con el fin de considerar la flexibilidad del receptor durante el *docking* y localizar diferentes sitios de unión alostéricos, crípticos (*i.e.* no accesibles en las estructuras conocidas), etc., **(c)** la caracterización energética de las interacciones proteína-ligando y **(d)** la identificación de mutaciones que aportan resistencia a fármacos conocidos permitiendo el avance de las terapias personalizadas.

Con esa finalidad, establecemos los siguientes objetivos específicos:

1. Desarrollar un protocolo automático de modelado por homología y aplicarlo en la reconstrucción tridimensional de las proteínas del centrosoma humano y otras proteínas de interés de las que no se dispone de su estructura tridimensional mediante técnicas experimentales. Asimismo, caracterizar proteínas desordenadas que no tienen una estructura única bien definida.

2. Estudiar el comportamiento dinámico, la flexibilidad y los cambios de conformación en proteínas de interés biológico mediante técnicas de dinámica molecular y modos normales torsionales.

3. Mejorar las predicciones teóricas de complejos moleculares incorporando las nuevas implementaciones en un protocolo de *docking* y cribado virtual y búsqueda de nuevos fármacos para la diana terapéutica OXA-24.

4. Implementación de un protocolo estadístico para el control de calidad, la identificación y la cuantificación de mutaciones puntuales con relevancia estructural provenientes de experimentos de *microarrays* de expresión.

4.- Trabajos de investigación

4.1.- Estructura y ausencia de estructura en proteínas del centrosoma humano

4.1.1. Introducción y aportación del autor

El centrosoma es un orgánulo molecular que regula procesos vitales de las células como las transiciones del ciclo celular, la respuesta al estrés, así como la división y diferenciación celular en metazoos. La estructura del centrosoma puede ser diferente según el tipo celular y el organismo. Está formado por un par de centriolos diferenciados con capacidad auto-replicativa, altamente estructurados y organizados formando típicamente nueve tripletes de microtúbulos que también pueden formar el cuerpo basal de cilios y flagelos. Es probable que los centriolos estuvieran ya presentes en el ancestro común de todos los eucariotas, sin embargo, no parecen ser imprescindibles durante la mitosis, la migración celular o el crecimiento axonal. Estos procesos requieren sin embargo del material pericentriolar (PCM) que los rodea, una matríz aparentemente carente de estructura definida que promueve la nucleación de los microtúbulos. Diversas enfermedades humanas han sido asociadas con anomalías en el centrosoma, tanto en relación al número de centriolos en casos de cáncer, como a nivel de mutantes que afectan al desarrollo cerebral. Conocer de un modo más detallado su estructura, la ausencia de estructura y su posible variabilidad conformacional sienta las bases para el desarrollo de nuevas terapias.

En un trabajo reciente llevado a cabo en nuestro laboratorio, se estudiaron las propiedades de 465 proteínas centrosomales humanas recogidas en la base de datos CentrosomeDB (http://centrosome.cnb.csic.es/, Nogales-Cadenas, Abascal, Díez-Pérez, Carazo, & Pascual-Montano, 2009) y se compararon con sus ortólogos en 6 especies animales y en levadura. Nido *et al.* (Nido, Méndez, Pascual-García, Abia, & Bastolla, 2012) observaron que las proteínas centrosomales de un organismo eran predichas como más largas, con más residuos *coiled-coil*, más sitios de fosforilación y más desordenadas que las proteínas control del mismo organismo, y que las regiones desordenadas crecen a lo largo de la evolución a través de grandes inserciones y proteínas de origen reciente. Además, la proporción de regiones desordenadas en el centrosoma y en proteínas control tiende a aumentar con el número de tipos celulares, lo cual establece una conexión entre la complejidad de los organismos y la complejidad molecular. Los autores proponen que la plasticidad estructural conferida por esas regiones desordenadas y los sitios de fosforilación podría jugar un papel importante en las propiedades mecánicas, así como en la regulación espacial y temporal del centrosoma.

La autora de esta tesis, usando la técnica de modelado por homología, ha llevado a cabo la construcción de un conjunto de modelos 3D compuesto por 361 proteínas cuya

localización centrosomal ha sido experimentalmente verificada (ver artículo). Para ello se desarrolló un protocolo completo y automático de modelado por homología con aplicación a conjuntos masivos de proteínas. Los pasos principales del dicho protocolo se resumen a continuación:

1.- Búsqueda y selección de un molde adecuado. Los moldes han sido identificados alineando las secuencias problema con una base de datos de perfiles *Hidden Markov Models* (HMM, véase sección 2.1.1.- Alineamiento de secuencias) construida a partir de las regiones con estructura resuelta definidas en el PDB, usando HHblits (Remmert et al., 2012) y conjuntos de datos del PDB agrupados con identidades de secuencia al 70%. Una vez realizada la búsqueda y obtenidos los alineamientos correspondientes, seleccionamos de manera iterativa los moldes que presentan un mayor número de residuos idénticos, utilizando sólo moldes con una ID mayor del 20% y menor del 95%, hasta cubrir el mayor porcentaje de la secuencia diana a reconstruir, permitiendo solapamiento entre los moldes encontrados si aportan más de 30 nuevos residuos.

2.- Obtención y ajustes del alineamiento molde-diana. Los alineamientos por pares obtenidos en el paso anterior se editan automáticamente descartándose regiones con grandes inserciones/delecciones entre molde y diana representadas como *gaps* en el alineamiento, así como residuos alineados aislados flanqueados por dichas zonas de *gaps*. En nuestro caso adoptamos una actitud conservadora donde consideramos que una zona de *gaps* no es modelable si supera las 6 posiciones.

3.- Construcción tridimensional de las proteínas diana. Llevamos a cabo las construcciones estructurales con MODELLER, obteniendo un conjunto de 20 modelos por fragmento modelable.

4.- Evaluación estructural. Seleccionamos el mejor modelo en base a su evaluación estructural y energética por métodos diferentes a los usados para su generación, que incluyen dos tipos de medidas: (1) estereoquímicas con el programa procheck (Laskowski, Macarthur, Moss, & Thornton, 1993) y los diagramas de *Ramachandran* y (2) energéticas con el *score* normalizado de DOPE (Fiser & Šali, 2003, ver sección 2.1.3.2- Evaluación energética).

5.- Refinado. Los modelos se han refinado minimizando la energía del campo de fuerza de AMBER, lo que permite eliminar choques estéricos y mejorar la red de HB. Observamos mejoras locales mediante la comparación de los perfiles por residuo DOPE, antes y después del

refinado. Los *scores* globales también indican, como tendencia general, que el proceso de refinado mejora las estructuras.

En resumen, los modelos 3Dobtenidos y las estructuras experimentales del PDB con ID > 95% engloban 361 fragmentos provenientes de 277 proteínas centrosomales y representan el 27,6% de los residuos del centrosoma humano. Las proteínas y regiones que no se han podido modelar corresponden en su mayoría (74%) a regiones predichas como desordenadas o *coiled-coil.* Los modelos 3D refinados, obtenidos de cada una de las proteínas o fragmentos a partir del mejor molde disponible en el PDB, y los moldes con ID > 95% han sido incorporadas en el artículo (Dos Santos et al., 2013) y puestos a disposición de la comunidad científica (http://ub.cbm.uam.es/centrosome/models/index.php) junto con otra información de interés como los alineamientos molde-diana, predicciones de desorden estructural y estructura secundaria y diversas tablas resumen que hacen de esta base de datos una herramienta útil y de referencia para aquellos investigadores que trabajen con proteínas centrosomales.

Artículo 1

OPEN ACCESS Freely available online

PLOS ONE

Structure and Non-Structure of Centrosomal Proteins

Helena G. Dos Santos[1☯], David Abia[1☯], Robert Janowski[2,3☯], Gulnahar Mortuza[4☯], Michela G. Bertero[5], Maïlys Boutin[2,3], Nayibe Guarín[2,3], Raúl Méndez-Giraldez[1], Alfonso Nuñez[1], Juan G. Pedrero[4], Pilar Redondo[4], María Sanz[5], Silvia Speroni[5], Florian Teichert[1], Marta Bruix[6], José M. Carazo[7], Cayetano Gonzalez[3,8], José Reina[3,8], José M. Valpuesta[7], Isabelle Vernos[5], Juan C. Zabala[9], Guillermo Montoya[4], Miquel Coll[2,3], Ugo Bastolla[1]*, Luis Serrano[5]*

1 Centro de Biología Molecular Severo Ochoa (CBMSO), CSIC-UAM, Madrid, Spain, **2** Institut de Biologia Molecular de Barcelona (IBMB), Baldiri Reixac 10–12, Barcelona, Spain, **3** Institute for Research in Biomedicine (IRB-Barcelona) Baldiri Reixac 10–12, Barcelona, Spain, **4** Centro Nacional de Investigación Oncológica (CNIO), Madrid, Spain, **5** Centre for Genomic Regulation (CRG), Barcelona, Spain, **6** Instituto de Química Física Rocasolano (IQFR), CSIC, Madrid, Spain, **7** Centro Nacional de Biotecnología (CNB),-CSIC, Madrid, Spain, **8** Institucio Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona, Spain, **9** IFIMAV-Universidad de Cantabria, Santander, Spain

Abstract

Here we perform a large-scale study of the structural properties and the expression of proteins that constitute the human Centrosome. Centrosomal proteins tend to be larger than generic human proteins (control set), since their genes contain in average more exons (20.3 versus 14.6). They are rich in predicted disordered regions, which cover 57% of their length, compared to 39% in the general human proteome. They also contain several regions that are dually predicted to be disordered and coiled-coil at the same time: 55 proteins (15%) contain disordered and coiled-coil fragments that cover more than 20% of their length. Helices prevail over strands in regions homologous to known structures (47% predicted helical residues against 17% predicted as strands), and even more in the whole centrosomal proteome (52% against 7%), while for control human proteins 34.5% of the residues are predicted as helical and 12.8% are predicted as strands. This difference is mainly due to residues predicted as disordered and helical (30% in centrosomal and 9.4% in control proteins), which may correspond to alpha-helix forming molecular recognition features (α-MoRFs). We performed expression assays for 120 full-length centrosomal proteins and 72 domain constructs that we have predicted to be globular. These full-length proteins are often insoluble: Only 39 out of 120 expressed proteins (32%) and 19 out of 72 domains (26%) were soluble. We built or retrieved structural models for 277 out of 361 human proteins whose centrosomal localization has been experimentally verified. We could not find any suitable structural template with more than 20% sequence identity for 84 centrosomal proteins (23%), for which around 74% of the residues are predicted to be disordered or coiled-coils. The three-dimensional models that we built are available at http://ub.cbm.uam.es/centrosome/models/index.php.

Citation: Dos Santos HG, Abia D, Janowski R, Mortuza G, Bertero MG, et al. (2013) Structure and Non-Structure of Centrosomal Proteins. PLoS ONE 8(5): e62633. doi:10.1371/journal.pone.0062633

Editor: Silvio C E Tosatto, Università di Padova, Italy

Received December 4, 2012; **Accepted** March 24, 2013; **Published** May 9, 2013

Funding: The authors gratefully acknowledge financial support from the Spanish Ministry of Science, grant CSD2006-00023, and from the Madrid Community, grant S2010/BMD-2305. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ubastolla@cbm.uam.es (UB); luis.serrano@crg.eu (LS)

☯ These authors contributed equally to this work.

Introduction

Since 1876, centrioles and centrosomes were shown to be involved in organizing fibrillar structures, including the spindle and mitotic asters within the cell as well as cilia and flagella in cells of many tissues [1]. The centrosome is a remarkable molecular machine capable of self-replication, whose core is constituted by two centrioles [2], highly structured macromolecular complexes typically consisting of nine microtubule-triplet-blades arranged in a cylinder, which also form the basal bodies required for the formation of cilia and flagella. Centrioles were probably present in the common ancestor of all eukaryotes [3] but, contrary to what was previously thought, they are not required for general mitosis, cell migration, and axonal growth [4]. Instead, these processes require pericentriolar material (PCM), a protein matrix that is the main constituent of the centrosome and apparently lacks higher order structure. Mutations in the centrosomes are related with several human diseases, most notably cancer [5] and abnormal brain development [6,7,8].

Recently, large-scale proteomic experiments have identified proteins localized in the human [9,10] and the fly centrosome [11]. Motivated by this study, Nogales-Cadenas et al. [12] retrieved from public databases, such as the Human Protein Reference Database (HPRD) [13], MiCroKit [14], Gene Ontology [15] and Ensembl [16] a large number of genes annotated as centrosomal from previous literature evidence as well as human orthologs of mouse centrosomal genes. A total of 465 likely centrosomal human genes, together with a rich set of biological annotations and derived information, were organized into a centralized resource named CentrosomeDB (http://centrosome.cnb.csic.es/).

The Centrosome 3D consortium is committed to analyze from a multidisciplinary point of view, including structural, cellular and

computational approaches, the physiology of this organelle. From the computational side, using bioinformatics predictions, we have observed that proteins forming the centrosome tend to be longer, more widely phosphorylated, and to contain a larger fraction of disordered [17] and coiled-coil [18] residues than control proteins of the same organism [19]. In particular, regions that are predicted to be simultaneously disordered and coiled-coil constitute a signature of centrosomal proteins. We have found that intrinsically disordered regions increased during the evolution of the centrosome through large insertions. Interestingly, insertions of disordered regions occurred at a faster rate along branches of the animal tree where the number of cell types of the organism experienced a large increase [19]. This observation suggests an intriguing relationship between the molecular complexity of the centrosome and the cellular complexity of the organism. As part of the effort of the Centrosome 3D consortium, we report here large-scale homology modeling and expression assays of the human proteins whose centrosomal localization has been experimentally demonstrated.

Methods

Selection of the Data Set

From the 465 genes in the CentrosomeDB database [12], we selected those genes whose evidence is based either in the Andersen *et al.* proteomic experiment with the human centrosome [9], or in the manually curated HPRD database [13], or in literature evidence obtained by text mining and manually verified, or it is supported by orthology with respect to experiments with the mouse centrosome [12] or the proteomic experiment with Drosophila [11]. We obtained 361 genes with solid evidence of centrosomal localization, which are listed in Table S1, discarding 104 genes from the Centrosome DB that do not fulfill the above criteria. We considered the longest isoform of each gene, whose sequences are reported in Table S2. 500 control human genes and the 1202 isoforms associated to them were randomly extracted from the Ensembl database. Their Ensembl codes and sequences are reported in Table S3.

Disorder, Coiled-coil and Secondary Structure Predictions

Disorder predictions were obtained with the DISOPRED2 [20], FoldIndex [21], IUPred [22] and DisEMBL [23] programs. In previous work, we tested that the first three algorithms have a large overlap with each other and produce qualitatively equivalent results. We present here predictions obtained with DISOPRED2, which was evaluated as the best among these predictors [24]. Predictions of coiled-coil residues were obtained with the NCOIL [25] and PCOILS [26] programs. Again, results are qualitatively equivalent and we show those obtained with NCOIL. Secondary structure was predicted with the PSIPRED program [27] and assigned with DSSP [28].

Template Selection

Suitable templates were obtained using Hidden Markov Models (HMM) [29] as implemented in the HHblits tool kit downloaded from http://toolkit.genzentrum.lmu.de/hhblits/. Namely, we used the HHblits software [30] to construct HMMs for the 28,020 representative protein chains in the PDB [31] clustered at 70 percent sequence identity. HHblits uses secondary structure to improve the constructed HMM. We searched each query sequence against these HMMs, keeping only highly significant matches (i.e. probability of being a true positive higher than 95%) with more than 30 residues and more than 20% sequence identity, which is considered as the minimum identity for obtaining reliable structural models. In order to retrieve structures with high sequence identity that are not chosen as representative structures to construct the HMMs, we searched with BLASTP [32] all 45,543 protein chains in the PDB clustered at 100 percent sequence identity, keeping only matches above 70% sequence identity. For each query sequence, we selected the templates yielding the maximum number of identical residues. Overlapping templates were kept if the less favored template contributes at least 30 new residues to be modeled.

Homology Modeling

For proteins with more than 20% and less than 95% sequence identity with a template structure in the PDB, structural models were built from the query-template alignment using the MODELLER program [33]. Model quality was assessed with the empirical energy function DOPE, implemented in MODELLER [34], with an empirical folding free energy function based on contact interactions [35] and with the program ProCheck [36], which checks the stereochemical quality of a protein structure, analyzing its overall and residue-by-residue geometry. Models were refined in order to avoid atomic clashes, allowing small relaxation through the following protocol: (1) System preparation: Hydrogen atoms and protons were added to the protein molecule using the program PDB2PQR [37] with the AMBER10 force-field [38] at pH 6.5; A water box of 10Å thickness was built around the protein with the program TLEAP [39] using the TIP3P model of water molecules [40]. Cl- and Na+ ions were added to neutralize when necessary. (2) Relaxation: The structure was refined by applying energy minimization followed by heating to 298 K, equilibration and cooling. No molecular dynamics per se was carried out due to the fact that, in many cases, models are quite small or too fragmented to be stable on their own. We employed NAMD 2.8 (Nano-scale Molecular Dynamics) [41] with the AMBER10 force field for the protein and the TIP3P model for water. First, the energy of the water molecules and ions was minimized keeping the protein fraction fixed. Second, the whole system was equilibrated at a constant temperature of 298 K, slowly reducing the constraints on the protein structure. Finally, the system was cooled, reducing the temperature from 298 K to 273 K with decrements of 1 K.

Cloning Centrosomal Genes

Cloning facilities at the CNIO in Madrid, at the IBMB and the CRG in Barcelona, and the company GenCust, produced clones of 138 centrosomal genes (Table S4), which are available upon request for academic use. Centrosomal proteins were cloned into pOPIN vectors using the In-Fusion™ PCR cloning method, a versatile ligation-independent cloning system engineered for high throughput screening [42]. Three different pOPIN vectors were used in this study: pOPINJ, pOPINS and pOPINM, utilizing the cleavable fusion tags His-GST, His SUMO and His-MBP, respectively. These vectors facilitate the expression of the cloned gene in *E. coli* or human cells and they can also be used to generate baculoviruses for insect cell infection.

Selection of Putative Globular Domains for Experimental Study

Putative globular domains were predicted for 208 selected centrosomal proteins of particular experimental interest by combining domain predictions through the SMART web server [43], disorder predictions [20], coiled-coil predictions [25] and sequence alignments of query centrosomal proteins against representative protein sequences in the PDB clustered at 50%

sequence identity. Sequences were parsed into predicted SMART domains, which were given higher priority, coiled-coil stretches (consecutive stretches of more than 20 predicted coiled-coil residues) and disordered stretches (consecutive stretches of more than 20 predicted disordered residues). Regions longer than 40 residues that were not classified as none of the above were aligned against representative sequences in the PDB using the program SABERTOOTH [44], which performs accurate alignments between distant homologs by aligning predicted structural profiles, is minimally influenced by sequence identity, and measures the significance of the alignment through a Z score.

Putative globular domains were identified if one of these conditions holds: (1) SMART finds a significant match with a known protein family over at least 30 residues; (2) SABERTOOTH finds a significant match with Z score >3 with a protein in the PDB over more than 40 residues, and the sum of disordered and coiled-coil residues is below 30%. Domains with more than 40% sequence identity with structures in the PDB were considered of little interest and discarded from further experimental study. We finally obtained 173 putative globular domains from which we selected for experimental studies 65 constructs belonging to 50 proteins (Table S5). Note that these putative domains selected for experimental study do not necessarily coincide with the structural domains predicted through our homology modeling procedure, since in this case we also used information about disorder, coiled-coil, and functional domains.

Expression of Full-length Proteins and Predicted Globular Domains

Full-length proteins corresponding to the longest isoform of centrosomal genes and selected globular domains were expressed and purified in three different labs (CNIO, IBMB and CRG), with common standardized protocols. For the expression tests the recombinant plasmids were used to transform *E. coli* B834(DE3) and Rosetta(DE3) pLysS cells. Cells were grown in LB media (2 ml) with appropriate antibiotics and induced at an OD600 of 0.8 with 0.3 mM IPTG. Two different temperatures were tested, harvesting cells after 3 h at 37°C and after 20 h at 20°C. In parallel, cells were also grown in auto-induction media and harvested after 20 h at 20°C, with shaking at 210rpm. Cells were lysed in standard buffers (50 m*M* Tris pH 8.5, 400 m*M* NaCl, 0.05% (*v*/*v*) Tween20) and overexpressed proteins purified either using NTA Ni spin columns or paramagnetic beads. Expression and solubility of the full-length proteins and the domains were checked by SDS-PAGE or Western Blot technique. The soluble proteins were confirmed by MALDI-MS analysis. Selected targets were also tested using baculovirus expression system.

Antibody Production

The functional facility of the consortium at the CRG produced 44 antibodies against 40 centrosomal proteins. Furthermore, 38 antibodies against 27 centrosomal proteins were produced by the company Eurogentech. These available antibodies are listed in Table S6. We also report in Table S7 antibodies against centrosomal proteins that were commercially available prior to our study.

Results

Predicted Disordered and Coiled-coil Fragments

More than 57% of the residues in human centrosomal proteins are predicted to be disordered in the native state. This fraction is significantly larger than for control human proteins, for which the fraction of predicted disordered residues, obtained with the same method as for centrosomal proteins, is 39%. These results hold for the longest isoforms. 72% of the centrosomal genes and 56% of the control genes have more than one isoform. In both cases, for these genes the shortest isoform is less disordered and less coiled-coil than the longest one, although this difference is only marginally significant (for instance, the disorder content is respectively 37.8% and 40.5% for the shortest and longest isoform of control genes, with a statistical error of 1%).

The distribution of the fraction of the longest isoform that is predicted to be disordered is shown in Fig. 1, distinguishing regions predicted to be disordered and coiled-coil and disordered and not coiled-coil. There is a significant positive propensity to predict a residue as coiled-coil if it is predicted to be disordered: 12.4% of the residues are predicted to be disordered and coiled-coil at the same time, compared with only 0.7% that are predicted to be coiled-coil but not disordered. Therefore, the propensity is $Prop(\text{coil\&disorder}) = \log(P(\text{coil\&disorder}) - \log(P(\text{coil})) - \log(P(\text{disorder})) = 0.51$. 55 proteins (15%) contain disordered and coiled-coil fragments that cover more than 20% of their length. The distribution of the fraction of the protein that is predicted as disordered and coiled-coil is shown in Fig. 1A. The same distribution for regions predicted to be disordered and not coiled-coil is shown in Fig. 1B. For control human proteins the fraction of residues predicted to be coiled-coil and disordered is much smaller (3.3%), and only 0.7% of the residues are predicted as coiled-coil and not disordered, resulting in a positive propensity between coiled-coil and disorder, $Prop(\text{coil\&disorder}) = 0.75$.

These residues predicted to be both disordered and coiled-coil may represent disordered regions that lack stable structure unless they interact with a binding partner, and take coiled-coil structure only upon binding. The fact that coiled-coil proteins can be disordered has been shown for several proteins, for example in the case of the Myc protein interacting with a competitor of its natural partner [45], and it is consistent with the finding that the sequence complexity of coiled-coil proteins is typically lower than for globular proteins [46].

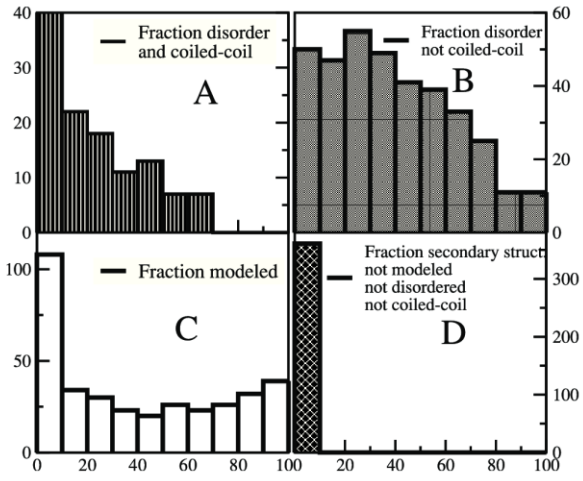


Figure 1. Fraction of protein length that is predicted to be disordered, coiled-coil, or modeled by homology. The plots represent the distribution of the percentage of protein length that has been (A) predicted to be disordered and coiled-coil at the same time; (B) Predicted to be disordered and not coiled-coil; (C) modeled; (D) Predicted to have regular secondary structure and not to be disordered neither coiled-coil, but not modeled.
doi:10.1371/journal.pone.0062633.g001

Homology Modeling

We modeled or retrieved the structures of 384 domains contained in 277 proteins. For 84 proteins (23%), no suitable templates were found. For these proteins, on the average 74% of the residues are predicted to be disordered in their native state and only 19 proteins are predicted to possess secondary structure in more than 30% of their residues. Globally, 27.6% of the residues were modeled. The histogram of the fraction of protein length that is modeled is shown in Fig. 1C, where one can see that, for most proteins, less than 50% of the length could be modeled. This lack of models is mainly due to structural disorder: 76% of the residues that were not modeled were predicted to be either disordered or coiled-coil, and the fraction of protein for which we could not build structural models, and which we predicted to possess secondary structure and to be neither disordered nor coiled-coil, was at most 10% (see Fig. 1D).

The length distribution of the 361 proteins (longest isoform of each centrosomal gene) is represented in Fig. 2A, where one can see that some proteins are extremely long. By contrast, the number of structural models built for each protein is almost in all cases smaller than 5 (Fig. 2B). The mean length of modeled fragments is 211 residues, ranging from 31 to 2922 residues (Fig. 3A). The distribution of sequence identity, plotted in Fig. 3B, is bimodal, with peaks at low and high identity: 99 fragments have less than 30% identity and 174 fragments have more than 90% identity with their PDB template. These templates with more than 95% identity were downloaded from the PDB, whereas lower identity templates were subject to the modeling procedure described in Methods.

The maximum length of gap regions modeled without a template was of 6 residues or fewer. Longer gaps cannot be reliably built, and they were left as unresolved structure. The DOPE energy, normalized so to transform it into a Z-score, is only slightly higher for the modeled sequence than for the template sequence (see Fig. 4A). Moderate energy structures could be slightly improved through the refinement protocol, but no improvement was achieved for high energy models, which may correspond to incorrect alignments. Similar results were obtained with the folding free energy function of Ref. [35], see Fig. 4B. Model quality was also assessed with ProCheck, which shows that the fraction of residues in disallowed regions of the Ramachandran plot increases not more than 4% from the template to the model, and that the number of residue pairs closer than 2.6Å (bad contacts) is on the average the same in the templates and the models. Based on these results, we conclude that the quality of templates and models is similar enough. However, some models had to be discarded by visual inspection, because they were very fragmented or presented too little secondary structure.

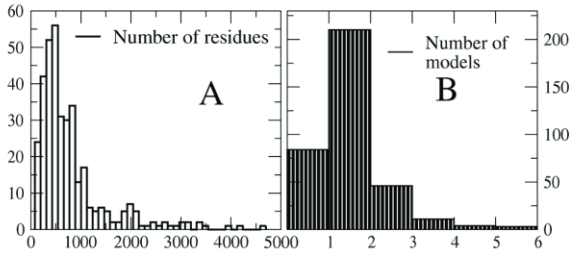


Figure 2. Number of residues and number of models for each protein. The plots represent the distribution of the number of residues of the longest isoform of centrosomal genes (A) and the number of structural models obtained for each protein (B).
doi:10.1371/journal.pone.0062633.g002

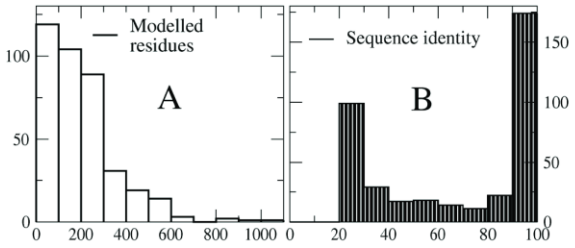


Figure 3. Summary of the structural models, either built by homology or retrieved from the PDB. The plots represent the distribution of the length (A) and sequence identity between query and template protein (B) for the 362 modeled fragments.
doi:10.1371/journal.pone.0062633.g003

Secondary Structure

Helices prevail over strands in modeled regions (44.8% against 16.8% as assigned by the program DSSP, which identifies secondary structures based on structural information, and 46.3 against 16.6% as predicted by the program PSIPRED, which only uses sequence information), and even more in the whole centrosomal proteome (52.0% against 7.0%, as predicted by PSIPRED), as expected due to the high incidence of coiled-coils and disordered loops. The frequency of predicted secondary structure classes is shown in Fig. 5.

For control human proteins, 34.5% of the residues are predicted as helical and 12.8% are predicted as strands, which means that centrosomal proteins are enriched in helical structures and depleted in beta strands. This difference between centrosomal and control proteins is mainly due to the large fraction of centrosomal residues predicted to be at the same time disordered and alpha-helical. In fact, of the residues predicted to be disordered in centrosomal proteins, 53% are predicted to be helical by PSIPRED and only 47% are predicted to be loops. In contrast, the fraction of predicted disordered residues of control human proteins that are predicted to be helical is only 25%, and the fraction predicted to be loop is 73%. As a consequence, 30.0% of the residues in centrosomal proteins are predicted to be disordered and helical, whereas this fraction is 9.7% in control proteins. This excess of residues predicted to be disordered and helical (20.3%) accounts for the difference between centrosomal and control proteins regarding helical residues (17.5%) and disordered residues (19.5%).

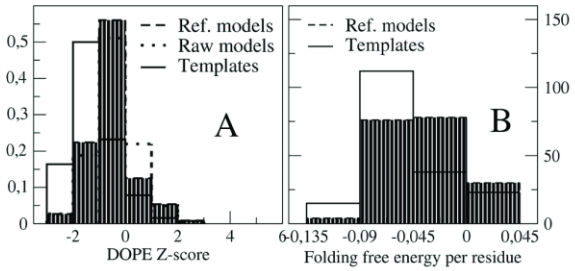


Figure 4. Empirical energy functions evaluated for each models and for the corresponding region of the template show that the predicted stability decrease is moderate.
doi:10.1371/journal.pone.0062633.g004

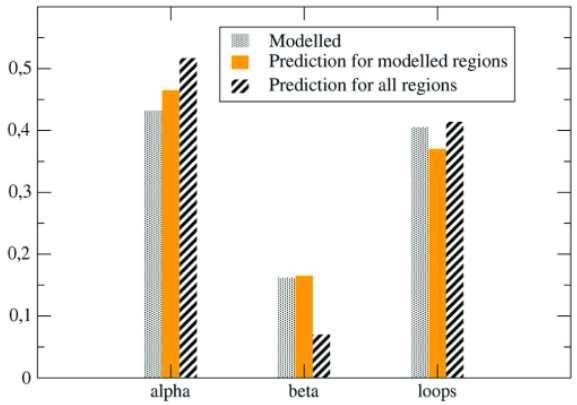


Figure 5. Frequency of the three main secondary structure classes for modeled residues (gray: DSSP of the template; yellow: PSIPRED prediction) and for all residues (pink). One can see that the set of all residues is strongly diminished in beta structures.
doi:10.1371/journal.pone.0062633.g005

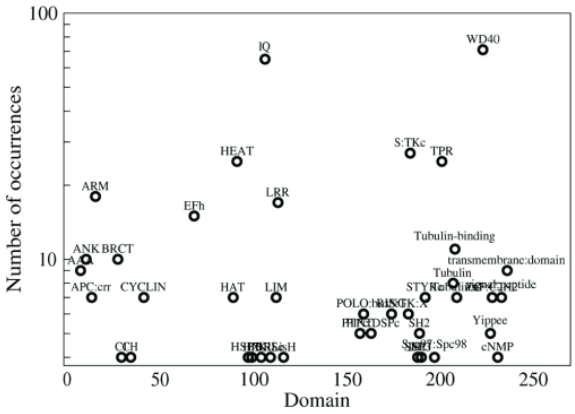


Figure 6. Number of occurrences of domains predicted by SMART. Only domains with more than 3 occurrences are shown.
doi:10.1371/journal.pone.0062633.g006

Online Database

We stored at the web site http://ub.cbm.uam.es/centrosome/models/index.php a database containing homology models and disorder and coiled-coil predictions for 361 human centrosomal proteins. For each protein there are available for online visualization and download disorder, coiled-coil and secondary structure predictions, homology models, and links to the UniProt (www.uniprot.org) and Centrosomedb [12] page. The number of modeled regions of each protein is indicated in parenthesis in the summary page, and for each model the user can download or visualize the three-dimensional structure superimposed with the template, the sequence alignment with the template, and the DOPE energy profile that identifies high energy regions that may be not well modeled. Models are linked to the PDB page of the template. The full set of models (only structures with less than 95% sequence identity) can be downloaded from the url http://ub.cbm.uam.es/centrosome/models/models_coordinates_95.tgz.

Modularity

Centrosomal proteins are highly modular. Besides coiled-coil regions, by far the most common structural motif, SMART identifies 719 evolutionary domains that belong to 239 types. The most frequent domains are the WD40 domain (71 occurrences), the IQ domain (65), the Serine Threonine Kinase domain (27), the TPR and HEAT domains (25), the ARM (18), LRR (17), EFh (15) and Tubuline binding (11) motifs. The number of occurrences for each domain in centrosomal proteins is represented in Fig. 6. Several modeled regions are constituted by multiple identical domains.

Protein-protein Interactions

We retrieved from the public databases DIP [47], MINT [48], INTACT [49], HPRD [50] and BOIGRID [51] the experimentally known protein-protein interactions of proteins in the human centrosome. We found 354 known interactions involving 167 of the 361 proteins. The average degree is 3.65 and the clustering coefficient is only 0.10. This low clustering coefficient suggests that the network is incomplete. We represent the network with the Cytoscape software [52] and we plot it in Fig. 7, representing in color code the betweeness centrality of each protein, a graph-theoretical measure of how central in the network is a node, which measures the number of shortest paths connecting any pair of nodes that pass through the given node [53]. Below, we list the most central proteins. We indicate in the brackets after the protein name the number of interaction partners and the fraction of the protein that is predicted to be disordered: TP53 (27, 46.5%), BRCA1 (20, 71.3%), YWHAG (13, 14.6%), APC (13, 84%), TUBG1 (11, 2.4%), DCTN1 (10, 94.5%), PIK3R1 (10, 30.6%), PLK1 (8, 31.4%), PAFAH1B1 (7, 5.2%). One can see that the mean disorder of central proteins is not very different from the average disorder of centrosomal proteins, but some of the central proteins, such as DCTN1, APC and BRCA1, are extremely disordered.

Full-length Proteins Expression and Purification

The length distribution of the 361 longest isoform of each centrosomal gene, represented in Fig. 2A, shows that the most populated length bin is for proteins slightly shorter than 500 residues, and a long tail of proteins longer than 1000 residues is present. The average length is 796±41 residues, which is significantly longer than the average length of 599±31 residues for control human proteins. One of us and coworkers observed in a previous work [19] that this difference is due to the fact that exons in human centrosomal genes are more numerous (20.3 versus 14.6) than exons in control human genes.

Out of the total number of 138 available clones of centrosomal genes (Table S4), 120 were successfully cloned in pOPIN expression vectors. Of these 120 full-length clones, 71 were subjected to high-throughput expression and purification methods, resulting in 24 soluble proteins (34%, Table S4), and 5 full-length proteins with unknown structure were expressed and purified to continue with further structural studies. The remaining 49 full-length clones were expressed and purified following a medium-low throughput method. 34 of these clones showed over-expression but only 15 were soluble and subjected to small-scale purification (31%, Table S4). Combining the two methods, only 39 out of 120 full-length proteins (32.5%) were soluble or partly soluble, in the sense that we found them both in the soluble fraction and in the pellet fraction.

Overall, centrosomal proteins were extremely tricky to handle and often the over-expression and solubility were highly sensitive to many external factors including the growth media, *E.coli* expression strains, the temperature and the induction time. Some proteins were soluble 3 hours after induction, however, by

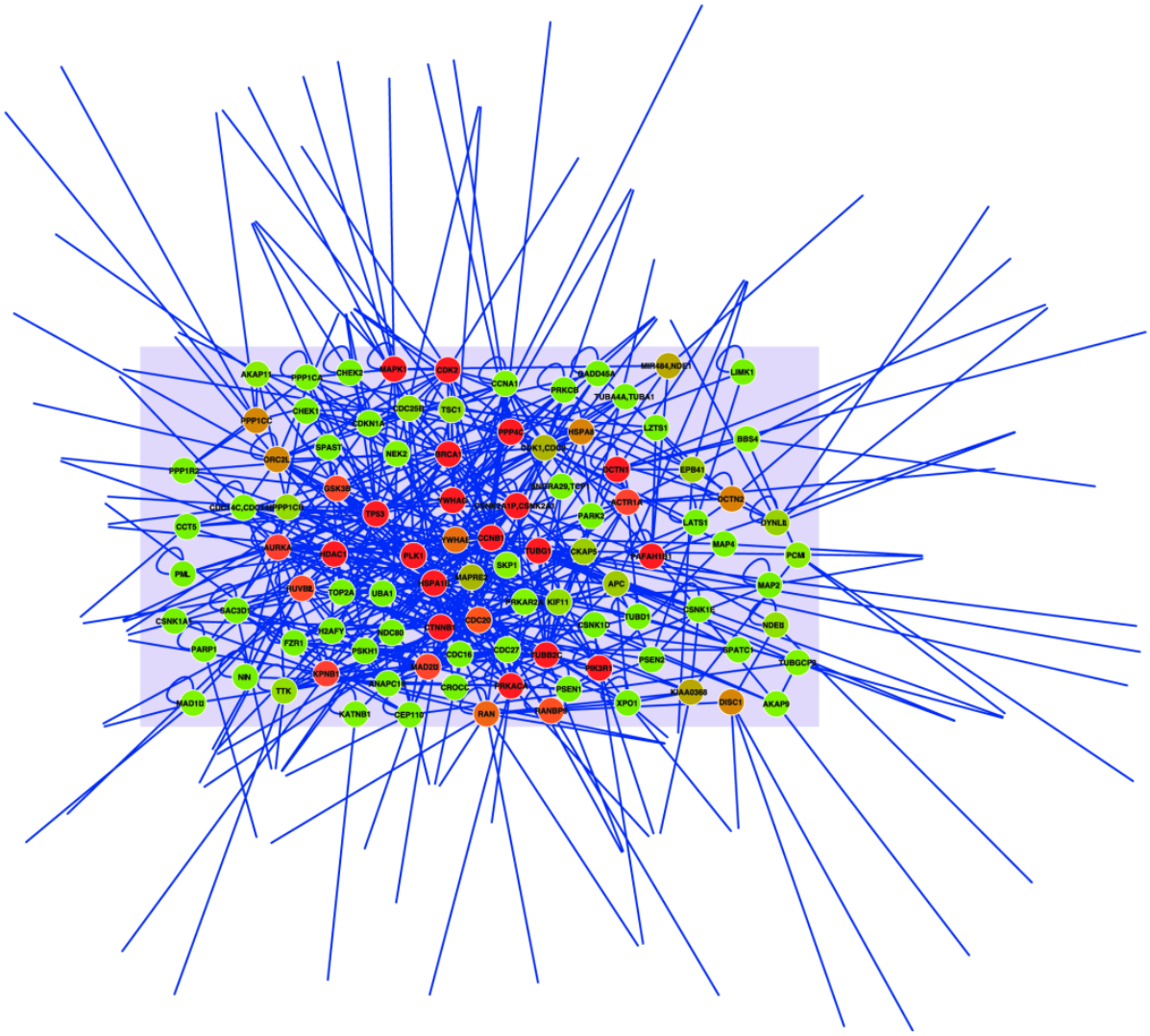


Figure 7. Protein-protein interaction networks for interactions experimentally observed for human centrosomal proteins. The color code represents betweeness centrality, a graph theoretic measure of the centrality of a node in a network, red representing the most central node. doi:10.1371/journal.pone.0062633.g007

decreasing the temperature and increasing the induction time of the same culture, they showed increased over-expression and became insoluble. Furthermore, solubility was compromised in several cases once the tag was removed indicating possible folding problems.

Domain Expression and Purification

Of the 173 domain constructs designed on the basis of the bioinformatics analysis and having unknown structure (see Methods and above), 44 domains were cloned in pOPINJ vectors and 22 in pOPINF and pOPINS vectors. All of them were expressed in high throughput conditions. Of these, only 14 (21%) were found to be soluble under the conditions tested and 5 were purified at high enough concentration for crystallization screenings (Table S5). Six additional domains were expressed in low-throughput conditions in CNIO, and 5 of them were found to be soluble and 4 were purified (see Table S5). Overall, we cloned 72 domains and found that 19 of them (26%) were soluble and 9 could be purified.

Discussion and Conclusions

Proteins in the centrosome tend to be long, modular, disordered and coiled-coil, significantly more than control proteins of the same organism. They are formed by a large number of exons, mostly corresponding to disordered regions, coiled-coils, or short domains such as the WD40 repeat, the IQ repeat and the HEAT repeat.

Centrosomal proteins are difficult to express: only 39 of the 120 full-length proteins in our expression trials were soluble (32.5%). Isoform length and disorder content can impact solubility. We took into account the disorder and coiled-coil content for predicting putative globular domains through a bioinformatics

analysis. We cloned and expressed these putative domains, but we obtained a similarly low success rate: only 19 out of 72 cloned constructs resulted in soluble proteins (26%). However, when domain prediction was coupled with low-throughput expression, the success rate greatly increased: 5 out of 6 domains cloned in these conditions were found to be soluble.

Experimental structures in the PDB or the structural models that we built through homology cover 27.6% of the length of centrosomal proteins. These modeled regions distribute quite unevenly in the predicted ordered and disordered regions. In regions predicted to be neither disordered nor coiled-coil, which represent 42.2% of centrosomal proteins, we could model 57.2% of the residues, whereas in regions predicted to be disordered or coiled-coil we could model only 5.4% of the residues. For 17.7% of the residues predicted to be in globular regions we could not find any suitable template, which demands further structural studies of globular domains in centrosomal proteins.

The main characteristics of centrosomal proteins are the numerous disorder and coiled-coil regions that make them extremely flexible and able to interact with many partners, forming intertwined coiled-coils. Interestingly, centrosomal proteins contain 30% of residues that are predicted to be disordered by DISOPRED and helical by PSIPRED, whereas this fraction is only 9.4% in control proteins. This difference accounts for most of the difference between centrosomal and control proteins concerning disordered residues (57.2% against 39%) and helical residues (52.0% against 34.5%). These regions are reminiscent of the proposed alpha-helix forming molecular recognition features (α-MoRFs), structural elements that mediate the binding events of initially disordered elements [54]. The role of intrinsically disordered protein regions in the interactions of centrosomal proteins has been experimentally demonstrated in a few cases. For instance, one of us and coworkers recently studied the N-terminal domain of the centrosomal protein TBCC that is involved in bipolar spindle formation. The TBCC-Nterm adopts a spectrin-like fold topology, and remarkably its 30-residue N-terminal fragment remains flexible and completely disordered in solution. The interaction of TBCC-Nterm with tubulin involves this unstructured region, which has been suggested to acquire structure upon interaction [55].

This structural complexity of centrosomal proteins and protein-protein interactions suggests that the centrosome will remain an important subject of structural investigation, which will probably require new experimental techniques.

Supporting Information

Table S1 List of the 361 genes with solid evidence of centrosomal localization considered in this study.
(DOC)

Table S2 Sequence of the longest isoform of each centrosomal gene.
(FASTA)

Table S3 Sequences of the 1202 isoforms associated to 500 control human genes that were randomly extracted from the Ensembl database.
(FASTA)

Table S4 List of the clones of 138 full-length centrosomal genes produced in this study, and their experimental characterization.
(DOC)

Table S5 Domain constructs selected for experimental studies and their characterization.
(DOC)

Table S6 Antibodies against centrosomal proteins produced in this study.
(XLS)

Table S7 Commercially available antibodies against centrosomal proteins retrieved through bioinformatics analysis.
(XLS)

Author Contributions

Conceived and designed the experiments: M. Bruix JMC CG JMV IV JCZ G. Montoya MC UB LS. Performed the experiments: HGDS DA RJ G. Mortuza MGB M. Boutin NG JGP PR MS SS. Analyzed the data: M. Bruix JMC CG JMV IV JCZ G. Montoya MC UB LS. Contributed reagents/materials/analysis tools: RM AN FT JR. Wrote the paper: UB LS G. Montoya MC G. Mortuza M. Bruix JCZ.

References

1. Nigg EA, Raff JW (2009) Centrioles, centrosomes, and cilia in health and disease. Cell 139: 663–678.
2. Ou Y, Zhang M, Rattner JB (2004) The centrosome: the centriole-PCM coalition. Cell Motil Cytoskel 57: 1–7.
3. Marshall WF (2009) Centriole evolution. Curr Opin Cell Biol 21: 14–19.
4. Hatch E, Stearns T (2010) The life cycles of centrioles. Cold Spring Harb Symp Quant Biol. 75: 425–31.
5. Boveri T (1914) Zur frage der entstehung maligner tumoren, Gustav Fischer, Jena. Translated and annotated by Henry Harris J (2008) Concerning the origin of malignant tumours. Cell Sci 121(1): 1–84.
6. Bettencourt-Dias M, Glover DM (2007) Centrosome biogenesis and function: centrosomics brings new understanding. Nat Rev Mol Cell Biol 8: 451–463.
7. Bond J, Woods CG (2006) Cytoskeletal genes regulating brain size. Curr Opin Cell Biol 18: 95–101.
8. Gerdes JM, Davis EE, Katsanis N (2009) The vertebrate primary cilium in development, homeostasis, and disease. Cell 137: 32–45.
9. Andersen JS, Wilkinson CJ, Mayor T, Mortensen P, Nigg EA, et al. (2003) Proteomic characterization of the human centrosome by protein correlation profiling. Nature 426: 570–4.
10. Jakobsen L, Vanselow K, Skogs M, Toyoda Y, Lundberg E, et al. (2011) Novel asymmetrically localizing components of human centrosomes identified by complementary proteomics methods. EMBO J. 30: 1520–35.
11. Müller H, Schmidt D, Steinbrink S, Mirgorodskaya E, Lehmann V, et al. (2010) Proteomic and functional analysis of the mitotic Drosophila centrosome. EMBO J. 29: 3344–57.
12. Nogales-Cadenas R, Abascal F, Diez-Pérez J, Carazo JM, Pascual-Montano A (2009) CentrosomeDB: a human centrosomal proteins database. Nucleic Acids Res 37: D175–80.
13. Mishra G, Suresh M, Kumaran K, Kannabiran N, Suresh S, et al. (2006) Human protein reference database–2006 update. Nucleic Acids Res 34: D411–4.
14. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, et al. (2008) Ensembl 2008. Nucleic Acids Res 36: D707–14.
15. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. Nat. Genet. 25: 25–9.
16. Ren J, Liu Z, Gao X, Jin C, Ye M, et al. (2010) MiCroKit 3.0: an integrated database of midbody, centrosome and kinetochore. Nucleic Acids Res 38: D155–D160.
17. Uversky VN, Dunker AK (2010) Understanding protein non-folding. Biochim Biophys Acta 1804: 1231–1264.
18. Lupas AN, Gruber M (2005) The structure of alpha-helical coiled coils. Adv Protein Chem 70: 37–78.
19. Nido GS, Méndez R, Pascual-García A, Abia D, Bastolla U (2012) Protein disorder in the centrosome correlates with complexity in cell types number. Mol Biosyst. 8: 353–67.
20. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337: 635–45.
21. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O (2005) FoldIndex(C): a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21: 3435–3438.
22. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21: 3433–3434.
23. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, et al. (2003) Protein disorder prediction: implications for structural proteomics. Structure 11: 1453–9.

24. Sirota FL, Ooi H, Gattermayer T, Schneider G. Eisenhaber F, et al. (2010) Parametrization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset. BMC Genomics (Suppl 1): S15.
25. Lupas A, Van Dyke M, Stock J (1991) Predicting coiled coils from protein sequences. Science 252: 1162–1164.
26. Gruber M, Söding J, Lupas AN (2006) Comparative analysis of coiled-coil prediction methods. J Struct Biol. 155: 140–5.
27. Buchan DW, Ward SM, Lobley AE, Nugent TC, Bryson K, et al. (2010) Protein annotation and modelling servers at University College London. Nucl. Acids Res. 38 Suppl, W563–W568.
28. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–637.
29. Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis, Cambridge Univ. Press.
30. Remmert M, Biegert A, Hauser A, Söding J (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods 9: 173–175.
31. The evolution of the Protein Data Bank (part of special essay collection on the 40th anniversary of the PDB) (2011) Nature Structural & Molecular Biology 18: 1310. Available: http://www.rcsb.org.
32. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.
33. Eswar N, Eramian D, Webb B, Shen MY, Sali A (2008) Protein structure modeling with MODELLER. Methods Mol Biol. 426: 145–59.
34. Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Prot. Sci. 15: 2507–24.
35. Bastolla U, Farwer J, Knapp EW, Vendruscolo M (2001) How to guarantee optimal stability tomost proteins in the Protein Data Bank. Proteins 44: 79–96.
36. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993). PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 26: 283–291.
37. Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, et al. (2007) PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucl. Ac. Res. 35 (suppl 2): W522–W525.
38. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, et al. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 65: 712–25.
39. Case DA, Cheatham TE 3rd, Darden T, Gohlke H, Luo R, et al. (2005) The Amber biomolecular simulation programs. J. Comp. Chem. 26: 1668–88.
40. Jorgensen W, Chandrasekhar J, Madura J, Impey R, Klein M (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79: 926–935.
41. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, et al. (2005) Scalable molecular dynamics with NAMD. J. Comp. Chem. 26: 1781–1802.
42. Berrow NS, Alderton D, Sainsbury S, Nettleship J, Assenberg R, et al. (2007) A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. Nucl Ac Res 35: e45.
43. Letunic I., Doerks T, Bork P (2010) SMART 7: recent updates to the protein domain annotation resource. Nucl Ac Res 40 (D1): D302–D305. Available: http://smart.embl-heidelberg.de/.
44. Teichert F, Minning J, Bastolla U, Porto M (2010) High quality protein sequence alignment by combining structural profile prediction and profile alignment using SABERTOOTH. BMC Bioinformatics. 11: 251.
45. Follis AV, Hammoudeh DI, Wang HB, Prochownik EV, Metallo SJ (2008) Structural rationale for the coupled binding and unfolding of the c-Myc oncoprotein by small molecules. Chem Biol, 15: 1149–1155.
46. Romero P, Obradovic Z, Dunker AK (1999) Folding minimal sequences: the lower bound for sequence complexity of globular proteins. FEBS Letters 462: 363–367.
47. Xenarios I, Salwínski L, Duan XJ, Higney P, Kim SM, et al. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucl. Ac. Res. 30: 303–5.
48. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, et al. (2002) MINT: a Molecular INTeraction database. FEBS Lett. 513: 135–40.
49. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, et al. (2004) IntAct: an open source molecular interaction database. Nucl Ac Res 32(Database issue): D452–5.
50. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009) Human Protein Reference Database–2009 update. Nucl Ac Res 37(Database issue): D767–72.
51. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, et al. (2006) BioGRID: a general repository for interaction datasets. Nucl Ac Res.34(Database issue): D535–9.
52. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, et al. (2007) Integration of biological networks and gene expression data using Cytoscape. Nature Protocols 2: 2366–2382. Available: http://www.cytoscape.org/.
53. Newman MEJ (2010) Networks: An Introduction. Oxford, UK: Oxford University Press.
54. Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, et al. (2005) Coupled folding and binding with alpha-helical forming molecular recognition elements (MoRFs). Biochemistry 44: 12454–12470.
55. Garcia-Mayoral MF, Castaño R, Fanarraga ML, Zabala JC, Rico M, et al. (2011) PLoS One. 6: e25912.

4.2.- Generación y evaluación de estructuras tridimensionales a partir de secuencias simuladas bajo diferentes modelos evolutivos de proteínas

4.2.1.- Introducción y aportación del autor

La simulación computacional trata de mimetizar procesos que ocurren en el mundo real permitiendo el estudio de sistemas complejos, testar hipótesis o incluso analizar datos reales. En el área de la evolución molecular es de interés simular la evolución de secuencias de ADN y proteínas teniendo en cuenta los procesos a diferentes niveles: **(1)** a nivel poblacional (por ejemplo demografía y migración), **(2)** a nivel molecular (por ejemplo mutación y recombinación) y **(3)** la interacción de ambos niveles (por ejemplo selección natural).

Por otro lado, los métodos probabilísticos de reconstrucción filogenética basados en evolución molecular (en nuestro caso *maximum likelihood*) requieren de modelos matemáticos sencillos del proceso de sustitución de los aminoácidos de las proteínas en función del tiempo. La mayoría de estos modelos asumen que cada residuo de la proteína evoluciona independientemente de los otros y usan matrices de las tasas de sustitución entre aminoácidos estimadas a partir de datos empíricos. Sin embargo, la hipótesis de independencia entre posiciones no es realista porque éstas están acopladas de manera global a través de la estabilidad del estado nativo y de la dinámica de la proteína. Por tanto la sustitución en una posición, al modificar la estabilidad, modifica a su vez la tasa de sustitución en otras posiciones. Estos efectos se pueden simular mediante modelos que representan el proceso de selección natural adoptando como objetivo la estabilidad de plegamiento (*fitness*) de la proteína, que se establece como criterio para calcular la probabilidad de que una mutación se fije en la población (sustitución) en función de su efecto predicho sobre la estabilidad.

El presente trabajo surge de la necesidad de incorporar modelos basados en estabilidad, como aquellos desarrollados desde varios años en nuestro laboratorio, en un programa que simule la evolución de las proteínas a lo largo de un árbol filogenético. Estas simulaciones se pueden aplicar para testar hipótesis, estudiar el efecto de unos parámetros evolutivos sobre otros (por ejemplo, cómo la recombinación puede afectar a la estabilidad de las proteínas resultantes) o incluso estimar parámetros evolutivos.

En el artículo (Arenas, Dos Santos, Posada, & Bastolla, 2013) presentamos el simulador evolutivo *ProteinEvolver*, un programa que implementa el modelo SCS (*Structurally Constrained Substitutions*) desarrollado en nuestro laboratorio, que basa la probabilidad de

una sustitución en la correspondiente variación de la estabilidad estructural del plegamiento (*fold*) de la proteína, es decir la *fitness* de un mutante depende de la estabilidad del estado nativo con respecto a su desplegamiento (*unfolding*) y/o plegamientos incorrectos (*misfolding*). Dicho programa está disponible gratuitamente para la comunidad científica a través de la web http://code.google.com/p/proteinevolver/.

En el trabajo, hemos comparado secuencias de proteínas generadas con el modelo SCS y con modelos de subsbtitución empíricos bajo el punto de vista de su capacidad de generar buenos modelos estructurales. Para ello se han usado 10 familias de proteínas obtenidas de la base de datos PFAM, simulando 200 secuencias sobre el árbol filogenético de cada familia para cada modelo de sustitución (por tanto, se han analizado un total de 200 secuencias x 2 modelos x 10 familias = 4000 secuencias simuladas). La aportación de la autora de esta tesis al presente trabajo ha consistido en generar 20 modelos 3D para cada secuencia simulada, evaluando su calidad mediante el *score* de DOPE, una función de energía independiente de la usada en las simulaciones.

Observamos que usando el modelo SCS el *score* de DOPE era en todos los casos más negativo, es decir las estructuras de las proteínas simuladas teniendo en cuenta la estabilidad estructural eran más estables que aquellas simuladas con el modelo empírico pero, como cabía esperar, a su vez menos estables que las estructuras cristalográficas usadas en las simulaciones. Los resultados nos muestran, además, que los modelos SCS generan distribuciones de aminoácidos cercanas a las observadas a familias de proteínas reales. Podemos concluir que los modelos evolutivos que consideran la estabilidad estructural de las proteínas constituyen una mejor aproximación a los procesos evolutivos reales y por tanto deben ser considerados para el estudio analítico de datos reales.

Artículo 2

Bioinformatics Advance Access published September 12, 2013

Original paper

Protein Evolution along Phylogenetic Histories under Structurally Constrained Substitution Models

Miguel Arenas[1,*], Helena G. Dos Santos[1], David Posada[2] and Ugo Bastolla[1]

[1]Centre for Molecular Biology "Severo Ochoa", Consejo Superior de Investigaciones Científicas (CSIC), Madrid, Spain.

[2]Department of Biochemistry, Genetics and Immunology, University of Vigo, Vigo, Spain.

Associate Editor: Dr. Janet Kelso

ABSTRACT

Motivation: Models of molecular evolution aim at describing the evolutionary processes at the molecular level. However, current models rarely incorporate information from protein structure. Conversely, structure-based models of protein evolution have not been commonly applied to simulate sequence evolution in a phylogenetic framework and they often ignore relevant evolutionary processes such as recombination. A simulation evolutionary framework that integrates substitution models that account for protein structure stability should be able to generate more realistic *in silico* evolved proteins for a variety of purposes.

Results: We developed a method to simulate protein evolution that combines models of protein folding stability, such that the fitness depends on the stability of the native state both with respect to unfolding and misfolding, with phylogenetic histories that can be either specified by the user or simulated with the coalescent under complex evolutionary scenarios including recombination, demographics and migration. We have implemented this framework in a computer program called *ProteinEvolver*. Remarkably, comparing these models with empirical amino acid replacement models, we found that the former produce amino acid distributions closer to distributions observed in real protein families, and proteins that are predicted to be more stable. Therefore, we conclude that evolutionary models that consider protein stability and realistic evolutionary histories constitute a better approximation of the real evolutionary process.

Availability: *ProteinEvolver* is written in C, can run in parallel, and is freely available from http://code.google.com/p/proteinevolver/.

Contact: marenas@cbm.uam.es, ubastolla@cbm.uam.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The simulation of molecular evolution is commonly used to mimic real world processes, allowing the study of complex systems that are analytically intractable or to understand the mechanisms by which an evolutionary process is modified (Arenas, 2012; Arenas, 2013; Hoban *et al.*, 2012). A key mechanism in the simulation of molecular sequences is the substitution or replacement process. Markov substitution/replacement models are commonly used in population genetics and molecular evolution to mimic evolutionary processes at the molecular level (see for a review Liò and Goldman, 1998). Nevertheless, most substitution models assume that sites evolve independently, and they cannot incorporate information on the structural and functional role of amino acids within proteins, which is determined by the interactions of different sites. It is known that these interactions may lead to non-independent evolution since the evolutionary rate at a site is influenced by substitutions at neighboring sites (e.g., Berard and Gueguen, 2012 and references therein). Moreover, the conformational diversity of proteins may also influence their molecular evolution (Javier Zea *et al.*, 2013; Juritz *et al.*, 2013). Models of evolution that incorporate structure are therefore of increasing importance (Anisimova and Liberles, 2007; Wilke, 2012). Recently, structurally constrained models of protein evolution have been introduced to represent folding stability of a target structure as a proxy for fitness (reviewed in Liberles *et al.*, 2012). These models have been studied assuming a neutral fitness landscape (e.g., Bastolla *et al.*, 2003; Bastolla *et al.*, 2006; Bastolla *et al.*, 1999; Parisi and Echave, 2001; Rastogi *et al.*, 2006; Taverna and Goldstein, 2002a; Taverna and Goldstein, 2002b) and as a function of population size (e.g., Goldstein, 2011b; Grahnen *et al.*, 2011; Mendez *et al.*, 2010). However, these models have been seldom used to obtain evolutionary insights from real data, probably because of the lack of widely available and easy-to-use software.

Furthermore, proteins simulated under a given substitution process might be unrealistic if common evolutionary processes are ignored. For example, recombination constitutes a fundamental evolutionary force at the molecular level (Posada *et al.*, 2002) that can affect the estimation of different evolutionary parameters, like molecular adaptation (by increasing the number of false positively selected sites) (Anisimova *et al.*, 2003; Arenas and Posada, 2010a; Kosakovsky Pond *et al.*, 2008), substitution rate (lack of molecular clock) (Schierup and Hein, 2000) or ancestral states (Arenas and Posada, 2010b) . Similarly, population genetics processes such as demographic changes, population structure and migration may influence evolutionary histories. For example, branches are short when the population size undergoes a bottleneck (e.g., Slatkin, 1996) and deme sizes in structured populations may influence the topology of genealogical trees since lineages are assumed to be in the same deme to coalesce (e.g., Neuhauser and Tavaré, 2001). These aspects could influence the number of substitutions and evolutionary trajectories generated in the simulations, which consequently influence the simulated data (Posada, 2001). In our view, it is important that a simulation framework allows

*To whom correspondence should be addressed.

reproducing and testing these evolutionary features in order to be able to address wider biological questions.

In this study, we have implemented structurally-constrained substitution models (hereafter, SCS models) that allow for site-dependent substitutions, under neutral and non-neutral fitness landscapes that depend on protein stability, in the freely available computer program ProteinEvolver, which is able to simulate the evolution of proteins and protein-coding genes along evolutionary histories such as phylogenetic trees or ancestral recombinant graphs (ARGs). These phylogenetic histories can be specified by the user or simulated through the coalescent with recombination, demographics and migration (see the reviews, Nordborg, 2000; Wakeley, 2008). Recently, Grahnen and Liberles introduced the computer program CASS that simulates protein sequence evolution under selection to fold into a specific conformation (Grahnen and Liberles, 2012). Our program differs from CASS in two main aspects: the representation of the protein structure, which influences the way in which we treat misfolding stability, and the possibility to represent a broad variety of demographic and evolutionary scenarios (for instance varying recombination rates and mutation). Whereas Grahnen and Liberles adopt all-atom representations of side chains, we adopt contact matrices, which are less precise but allow a statistical mechanical treatment of the ensemble of misfolded conformations that is computationally affordable (and actually very fast with our approximation) (e.g., Bastolla *et al.*, 2005a; Bastolla *et al.*, 2005b). Misfolding stability is important because, if only the unfolded state is considered, selection tends to artificially favor very hydrophobic sequences. Therefore, the structural approach taken by the two programs is complementary, and each of them may be suited to address different kinds of biological questions.

Through extensive simulations we compared the SCS models with commonly used empirical amino-acid substitution models using as a benchmark 10 well-known protein families. We found that sequences simulated under our SCS models produce amino acid distributions closer to the observed ones. Furthermore, the folding stability of the native state, assessed by building structural models by homology and predicting their stability with a method different from the one adopted in the simulations, is significantly larger for proteins simulated under the SCS model. We conclude that substitution models that incorporate protein structure information are better approximations to the real evolutionary process and may provide more meaningful evolutionary inferences than site-independent substitution models.

2 METHODS AND ALGORITHMS

We simulate protein evolution in two main steps. First, the genealogy is either specified by the user or it is simulated using the coalescent optionally modified with recombination, migration and demographics. Second, protein-coding genes and protein sequences are evolved along this genealogy under a given substitution or replacement model.

2.1 Simulation of genealogies

Genealogies are simulated according to the standard coalescent process (Kingman, 1982) modified with recombination (Hudson, 1983), demographics and migration (Hudson, 1998). Recombination can be either homogeneous or heterogeneous (recombination hotspots and coldspots) across the sequence following the algorithm developed by Wiuf and Posada (2003). Note that the simulation of recombination events leads to reticulate nodes and, consequently, to an ARG (Griffiths and Marjoram, 1997). Demographics include growth rate and demographic periods by following the algorithms implemented in Arenas and Posada (2007; 2010a). Gene flow among subpopulations can be specified under island, stepping-stone or continent-island migration models (e.g., Hudson, 1998). In addition, longitudinal samples or population/species trees, among other capabilities, can also be specified (see Table S1; supplementary material).

2.2 Evolution of proteins along phylogenies

After the phylogenetic history has been specified, a protein structure and a sequence are assigned to the root (the most recent common ancestor -MRCA-, or grand most recent common ancestor -GMRCA- in case of recombination). Then, the protein is evolved along the phylogeny going forward in time from the root to the tips (see, Yang, 2006) according to the SCS model, generating a protein for all internal and terminal nodes in the phylogeny.

Overall, for a given branch the SCS models perform five steps.

(i) The number of mutations or substitutions is computed considering the branch length (number of expected substitutions without considering structural constraints) and the length of the protein (number of amino acids).

(ii) A mutation is introduced according to the instantaneous rate matrix (the relative rates of change can be used to determine mutational sites).

(iii) The folding free energy of the mutated protein structure is computed.

(iv) The selective effect of the mutation is evaluated; it will be accepted or rejected depending on the fixation probability associated with the change of fitness (see next section).

(v) If the mutation is rejected, the process goes to "(ii)" and a new mutation is introduced. If the mutation is accepted, then it the mutation is fixed and it becomes a substitution.

The five-step process is repeated until the number of mutations or substitutions (i) is completed. In this way, fixed mutations (substitutions) result in differences among proteins from different evolutionary lineages.

2.2.1 Substitution models based on the stability of the protein structure

Evaluation of the structural stability of mutated proteins.

We evaluate the folding stability of a given mutated protein taking into account the stability against unfolding and against misfolding. Initially, we estimate the stability of the mutated sequence folded into the target structure at the simulation temperature using a contact-based free energy function. The contact matrix C_{ij} takes the value 1 if residues i and j are 'close' (<4.5Å) in space and 0 otherwise. This matrix has been shown sufficient to accurately reconstruct the three-dimensional structure of the protein (Vendruscolo *et al.*, 1997). We assume that the free energy of a protein with sequence A folded into the contact matrix C is given by the sum of its pairwise contact interactions:

$$E(A,C) = \sum_{ij} C_{ij} U(A_i, A_j) \qquad (1)$$

where $U(a,b)$ is the contact interaction matrix that expresses the free energy gained when amino acids a and b are brought into contact determined in (Bastolla *et al.*, 1999). For proteins that fold with two-state thermodynamics, i.e. for which only the native structure and the unfolded structure are thermodynamically important, stability against unfolding is defined as the free energy difference between the folded and the unfolded states, estimated as $\Delta G \sim E(A,Cnat)+sL$. Here *Cnat* is the native structure, L is the protein length, s is an entropic parameter and sL is the free energy of the unfolded state for proteins with two-states thermodynamics. We use $s = 0.074$, a value that was determined fitting the above equation to a set of 20 experimentally measured unfolding free energies, yielding a correlation coefficient r = 0.92. The correlation coefficient between the predicted and the observed stability effect of mutations is larger than 0.8 using only two fit parameters, which is comparable to state-of-the-art atomistic methods such as Fold-X (Guerois *et al.*, 2002).

Stability against unfolding is however not sufficient to characterize protein stability. We also have to check the stability against compact, incorrectly folded conformations of low energy that can act as kinetic traps in the folding process and, in many cases, result in pathological aggregations. Stability against misfolding is achieved by natural proteins by increasing the energy of key contacts that are frequently found in alternative structures, which is termed negative design (Berezovsky *et al.*, 2007; Minning *et al.*, 2013; Noivirt-Brik *et al.*, 2009) to distinguish it from the positive design that favors protein stability by strengthening native interactions. Therefore, stability against misfolding may be influenced by mutations at positions that are distant in the native structure. Stability against misfolded structures is difficult to estimate, and most models of protein evolution do not consider it despite its importance being increasingly recognized (Krishna *et al.*, 2004; Mendez *et al.*, 2010; Zheng *et al.*, 2013). Here we do consider the set of alternative compact matrices of L residues that can be obtained from non-redundant structures in the Protein Data Bank. This procedure, called *threading*, guarantees that the contact matrices fulfill physical constraints on chain connectivity, atomic repulsion, and hydrogen bonding (secondary structure), which are not enforced in the contact energy function. The free energy of this misfolded ensemble is often estimated with the Random Energy Model [REM, (Derrida, 1981)]:

$$G_{misfold} \approx \langle E(A,C)\rangle - \frac{\sigma^2}{2k_BT} - k_B T s_c L \qquad (2)$$

where $\langle E(A,C)\rangle$ is the mean and σ^2 is the variance of the energy of alternative compact structures (Goldstein, 2011a). This formula holds for temperatures above the freezing temperature at which the entropy of the misfolding ensemble vanishes. At lower temperatures the free energy maintains the same frozen value (Derrida, 1981). A recent study showed that the third moment of the energy cannot be neglected (Minning *et al.*, 2013), so that the free energy of the misfolded ensemble can be computed as,

$$G_{misfold} \approx \sum_{ij} \langle C_{ij}\rangle U_{ij} - \frac{1}{2k_BT}\sum_{ijkl} \langle C_{ij}C_{kl} - \langle C_{ij}\rangle\langle C_{kl}\rangle\rangle U_{ij}U_{kl} + \frac{1}{6(k_BT)^2}\sum_{ijklmn} \langle (C_{ij}-\langle C_{ij}\rangle)(C_{kl}-\langle C_{kl}\rangle)(C_{mn}-\langle C_{mn}\rangle)\rangle U_{ij}U_{kl}U_{mn} - k_BTs_cL \qquad (3)$$

where we denote with $U_{ij}=U(A_i,A_j)$ the contact free energy between residues i and j, and with $\langle C_{ij}\rangle$ the contact-specific mean value of the contact between the pair of residues i and j in a large set of compact protein structures of the same length L as the target structure.

In the present work, we have reduced considerably the computation time approximating the above free energy (Minning *et al.*, 2013) with one that only depends on pairs of residues,

$$G_{misfold} \approx A^{(1)}\langle U\rangle - \frac{A^{(2)}\langle U\rangle^2 + \sum_{ij} B_{ij}^{(1)} U_{ij}^2}{2k_BT} + \frac{A^{(3)}\langle U\rangle^3 + \langle U\rangle\sum_{ij} B_{ij}^{(2)} U_{ij}^2 + \sum_{ij} B_{ij}^{(3)} U_{ij}^3}{6(k_BT)^2} - k_BTs_uL \qquad (4)$$

The quantities $A^{(1)}\ A^{(2)}\ A^{(3)}\ B_{ij}^{(1)}\ B_{ij}^{(2)}\ B_{ij}^{(3)}$ only depend on the set of alternative contact matrices and on protein length L, and they are pre-computed before the simulation starts. In this way, we can evaluate how the misfolded free energy changes upon mutation performing only order L operations for computing $U(A_i = b, A_j) - U(A_i = a, A_j)$ when the residue at the mutated site i changes from amino-acid a to b. Thus, the stability of the native state is finally evaluated as the difference in free energy between the native, the unfolded and the misfolded states, $\Delta G = E(A,C_{nat}) - G_{misfold} - k_BTs_uL$.

The statistical properties of alternative contact matrices are computed from a large set of protein structures, distributed with the *ProteinEvolver* package, that can be modified by the user. Supplementary Figure S5 shows the histogram of the lengths of the alternative structures.

Note that, even if the two configurational entropies per residue S_u (unfolded ensemble) and S_c (misfolded ensemble) act additively, the free energy may not simply depend on their sum, since it is only S_c that determines the freezing temperature of the misfolded ensemble.

Relationship between protein stability and fitness.

Once we have defined protein stability, for modeling protein evolution we still have to define how protein stability influences fitness. Our program provides two alternatives. The simplest possibility is a neutral fitness landscape where the fitness is a binary variable and all proteins with stability above a given threshold, i.e. $\Delta G < \Delta G_{thr}$ are considered viable and equally fit, whereas all proteins below threshold are considered lethal and therefore discarded. We choose as threshold the folding free energy of the protein sequence A_0 in the Protein Data Bank $\Delta G_{thr} = \Delta G(A_0, C_{nat})$. This choice implies that the neutral SCS model is not sensitive to variations of the entropy parameters and it is little sensitive to variations in temperature.

Alternatively, we can consider a non-neutral scenario in which the probability of mutations being fixed depends on population size. In this case, there will be segregating variation in a population**.** Here, the fitness landscape is modelled in such a way that fitness is an increasing function of stability, and in particular it is proportional to the fraction of protein that is in the native state (Goldstein, 2011a),

$$f(A) = \frac{1}{1 + e^{\Delta G(A, C_{nat})/kT}} \qquad (5)$$

Note that the fitness landscape can be reduced to the neutral landscape in the low temperature limit, since in this limit the fitness tends to 1 if $\Delta G < 0$ and to zero if $\Delta G > 0$. This is a neutral landscape with $\Delta G_{thr} = 0$.

We then assume that the mutation rate is small and we model selection through the Moran's birth-death process (Ewens, 1979), which yields the fixation probability,

$$\Pi(ij) = \frac{1-\left(\frac{f_i}{f_i}\right)^a}{1-\left(\frac{f_i}{f_j}\right)^{2N}} \quad (6)$$

where f_i is the fitness of the wild-type, f_j is the fitness of the mutant, N is the effective population size and a = 2 or 1 for a haploid or diploid population, respectively. Given the probability of fixation, the succession of mutant fixations can be depicted as a Markov process, in which the genotype of the evolving lineage moves from one sequence to another one according to the mutation and fixation probabilities. Both the neutral and non-neutral scenarios are formally equivalent to a Monte Carlo process in statistical mechanics, as discussed by Sella and Hirsch (2005). The main difference between them is that in the neutral case, evolved proteins attain the minimum stability compatible with viability (which in this case is a parameter of the model), as discussed by Taverna and Goldstein (2002b), whereas in the non-neutral scenario stability increases with population size, and it also depends, in a non-trivial way, on the statistical properties of the mutation process (Mendez *et al.*, 2010). Therefore, neutral simulations depend on fewer parameters and they are more robust while non-neutral ones allow to explore more biological questions.

2.2.2 Simulation of the SCS model along an ancestral recombination graph

SCS models are site-dependent, so the simulation across an ARG is not straightforward, since recombination events put together in the same sequence sites that have been evolving independently along different lineages. In order to evolve the protein as a whole across the ARG, we adapted the algorithm developed by Arenas and Posada (2010a) to evolve codons "broken" by recombination (see Figure 1). The protein evolution occurs from the ancestral to the descendant nodes (see, Yang, 2006). However, if the evolutionary process reaches a recombinant node (Figure 1, nodes in grey), a protein is assigned to such a node (Figure 1, step 3), but at this point the evolutionary process continues along another path (Figure 1, step 4) because its parental recombinant node remains empty (without an assigned protein). This is forced to occur because in the parental recombinant node there is no information about the protein, since the evolutionary process did not reach it yet. Later, the evolutionary process reaches the parental recombinant node (Figure 1, step 5) and, at this point, there are entire proteins assigned to both recombinant nodes. Therefore, now there is a combination of the material according to the recombination breakpoint. This combination results in a new protein (Figure 1, step 6) that continues the evolutionary process along its descendant branch.

3 FIT OF THE SCP MODELS TO REAL PROTEIN FAMILIES

We studied 10 protein families in order to compare the performances of the integrated SCS models versus the empirical amino acid substitution models. Also, we estimated the temperature and entropic parameters that best reproduce the observed data.

We randomly selected 10 different protein families (Table 1) from the *Pfam* Database (http://pfam.sanger.ac.uk/) subject to two requirements: the *Pfam* seed alignment must possess at least 10 proteins and at least one representative structure included in the Protein Data Bank (PDB, http://www.rcsb.org/pdb/). For each protein family we downloaded the seed alignment and its phylogenetic tree from the *Pfam* database, and chose a representative structure from the PDB. Amino acid positions not included in the protein structure were trimmed from the alignment. The empirical amino acid replacement model that fitted best each alignment was estimated with *ProtTest* (Abascal *et al.*, 2005) (see Table 1).

For each protein family, we simulated 200 realizations of the evolutionary process under the best-fitted empirical amino acid substitution model along the *Pfam* tree. We also performed 200 realizations under the neutral SCS model and the fitness SCS model using the representative PDB structure. For both neutral and fitness SCS models we explored 27 combinations of the thermodynamic parameters temperature (T=1.75, 1.50, 1.30), configurational entropy per unfolded residue (s_u=0.025, 0.05, 0.075) and configurational entropy per misfolded residue (s_c=0.025, 0.05, 0.075), performing a total of 11,000 simulations. For the fitness SCS model we assumed an effective population size N=100 (Cruzan, 2001; Oostermeijer *et al.*, 1994). Note that all substitution models simulated an overall similar number of substitution events (Table S2; supplementary material) because they were applied along the same branch lengths.

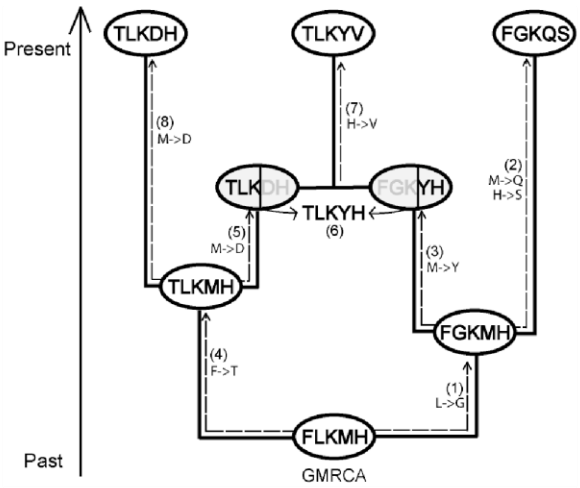


Fig. 1. An example of protein evolution along the ARG. White and grey circles correspond to coalescence and recombination parental nodes, respectively. (1) Starting from the GMRCA, the protein is evolved along branches according to the SCS substitution model and the branch lengths. (3) The process encounters a recombinant node and because its parental node has not been assigned to a protein yet, the evolutionary process continues towards other direction (4). (5) Later, the process encounters the parental recombinant node, and because the other parental has already been assigned to a protein, (6) it combines the two proteins according to the recombination breakpoint.

3.1 Analysis of amino acid distributions and inferences of optimal thermodynamic conditions

We compared the simulated amino acid distributions with those observed in the original data sets measuring their Kullback-Leibler divergence at each site *i*:

$$d_{KL,i} = \sum_{a=1}^{20} P_{\hat{i}}^{obs}(a)\left(\log\left(P_{\hat{i}}^{obs}(a)\right) - \log\left(P_{\hat{i}}^{sim}(a)\right)\right) \quad (7)$$

where *a* is any of the 20 amino-acids. We compute the weighted sum $D_{KL} = \sum_{i=1}^{L} w_i d_{KL,i}$ with weights w_i proportional to the number of aligned residues (excluding gaps) in column *i* of the alignment and summing up to one. The smaller the quantity D_{KL}, the closer the observed and simulated distributions are.

Remarkably, we found that sequences simulated under the neutral SCS model are always closer to the observed distribution than sequences simulated under the empirical substitution model (see Figures 2 and S1; supplementary material). Varying the temperature or the configurational entropy parameters did not affect the divergence of the neutral SCS simulations, as it could be expected since the difference between the estimated ΔG and the neutral threshold ΔG_{thr} is independent of the entropic parameter and it depends only weakly on temperature.

On the other hand, the divergence of the fitness SCS model clearly depended on the particular thermodynamic parameters (see Figures 2 and S2; supplementary material). The average agreement between predicted and observed site-specific amino acid distributions in the fitness SCS model reached a minimum for the combination of entropic and temperature parameters (see Figure S3, minimum in the left plot), where its value was similar to the one produced by the neutral model that used as threshold the free energy of the sequence in the PDB (Figure S3, right plot). These optimal thermodynamic parameters were achieved for $T(s_u+s_C)$=0.16. Interestingly, for some protein families, the fitness SCS model with optimal parameters was significantly less divergent from the observed distribution than the neutral SCS model (see Figure 2). However, for some protein families the fitness SCS model with the 'worst' parameter values had a similar divergence from the observed sequences as the empirical substitution models. These findings indicate that the neutral SCS model is more robust, and suggest that it should be used by default, whereas the fitness SCS model may be used for refinement with well calibrated thermodynamic parameters, as shown in Figure S3 (left). Comparisons between real sequence families, SCS simulations and simulations based on empirical amino acid substitution models are shown as sequence logos in Figure S6 (supplementary material).

3.2 Structural assessment of simulated proteins respect to the real proteins

We also assessed how much a simulated protein sequence fits a representative protein structure of its family by using homology modeling techniques (Marti-Renom *et al.*, 2000). For each protein family, 200 sequences simulated under the neutral SCS model and under the best-fit empirical site-independent substitution model, were modeled using the *Modeller s*oftware (Eswar *et al.*, 2006; Sali and Blundell, 1993). For each simulated sequence, 20 structural models were generated and they were assessed through their discrete optimized protein energy (DOPE) score (Shen and Sali, 2006), an effective energy function designed for selecting the best model built by *Modeller*. Note that this energetic score is independent from the one used in the SCS models. Then, we selected the sequence-structure pair with the lowest DOPE energy, whose sequence identity with the template is reported in Table S3 (supplementary material). We computed the DOPE energies for the experimentally known sequence-structure pair and for the best structural models of proteins simulated with the neutral SCS model and the best-fit empirical substitution model. Clearly, proteins simulated with the SCS model resulted in better sequence-structure pairs than proteins simulated with the empirical amino acid substitution model (see Figures 3 and S4; supplementary material). This result is not surprising, since we observed that the DOPE score was correlated with the contact energy of the native structure for proteins simulated under the SCS model. However, the two empirical energy functions were derived under different assumptions, therefore the DOPE score may be regarded as another confirmation of the quality of our models. Of course it is expected that models based on substitution matrices, which do not take into account the structure, produce proteins that are less stable than proteins simulated under SCS conditions. Nevertheless the explicit proof of this expectation is a necessary test to assess the necessity of SCS approaches.

4 SOFTWARE IMPLEMENTATION

We implemented the algorithms described above in the program *ProteinEvolver*. The full list of capabilities of *ProteinEvolver* is shown in the Table S1. *ProteinEvolver* is written in C, it can be run in parallel using MPI and it is freely available under the GNU GPL license from http://code.google.com/p/proteinevolver/. The package includes executables, source code, a detailed documentation and several practical examples (including the files and settings to mimic the evolution of the real proteins described and analyzed in the previous section).

Table 1. Protein families collected from the *Pfam* database

Entry	Protein family	Pfam code	Sample size	Uniprot entry	PDB code	Protein length	Best-fit amino acid model
1	Phototactive yellow proteins	PF00989	49	PYP_HALHA	2PHY	125	WAG +G +F
2	Triosephosphate isomerases	PF00121	56	TPIS_TRYBB	1TTI	243	RtREV +I +G +F
3	Rubredoxins	PF00301	43	RUBR2_PSEOL	1R0F	54	WAG +I +G
4	Kinesins	PF00225	87	KAR3_YEAST	3KAR	346	LG +I +G +F
5	Phage lysozymes	PF00959	18	LYS_BPT4	1OV5	164	Blosum62 +G +F
6	Ferredoxins	PF05996	62	PCYA_SYNY3	3NB8	248	WAG +I +G
7	DNA ligases	PF13298	136	B1L4V6_KORCO	3P4H	118	WAG +G
8	Heat shock proteins	PF00012	33	DNAK_ECOLI	2KHO	605	RtREV +G +F
9	Oxysterol-binding proteins	PF01237	153	KES1_YEAST	1ZHT	438	LG +I +G +F
10	Retroviral aspartil proteases	PF00077	50	POL_FIVPE	3OGQ	116	RtREV +I +G

For each family, the table indicates the *Pfam* code, sample size, *UniProt* entry for a protein sequence with a PDB structure, the PDB code, number of amino acids and the empirical amino acid substitution model that better fits the dataset [+G indicates variable substitution rate across sites according to a gamma distribution, +I indicates a proportion of invariable sites and, +F indicates amino acid frequencies].

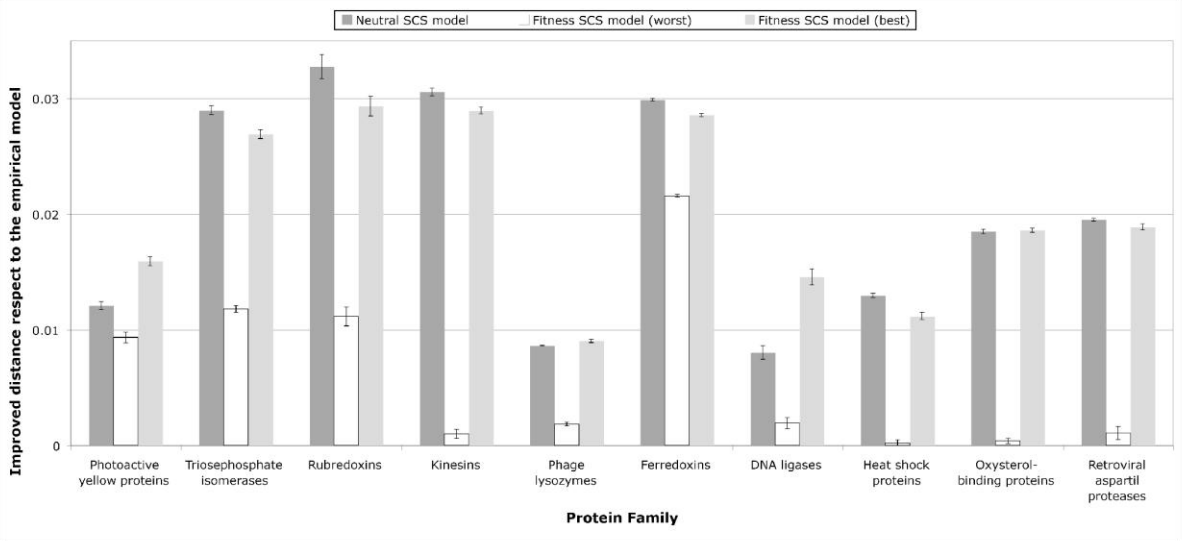


Fig. 2. Improvement of the Kullback-Leibler distance to the real protein alignments of the simulated alignments by the neutral and fitness SCS models with respect to the empirical amino acid substitution model. The "y" axis indicates decline of the distance of the neutral and fitness (best and worst conditions) SCS models respect to the distance of the empirical model. Note that the neutral SCS model was overall more robust than the fitness SCS model under different thermodynamic conditions (see Figures S1A and S1B). On the other hand, the fitness SCS model under the best conditions (see Figures S2A, S2B and S3, left plot) could improve the neutral model in half of protein families, however the worst conditions may lead to results without any improvement respect to the empirical model.

5 DISCUSSION

During protein evolution, interactions within the protein structure lead to correlated evolution, since the rate at which a site experiences change is influenced by replacements at neighboring sites. To adequately model these correlations, we developed the simulation framework *ProteinEvolver* that integrates structure-based models of protein evolution and evolutionary histories that can be simulated under diverse evolutionary scenarios such as recombination (including hotspots and coldspots), migration and demographics. Importantly, our SCS models consider both the stability against unfolding and the stability against misfolding, which is difficult to estimate since it requires the use of a set of alternative conformations, and it is frequently neglected in simulations of protein evolution despite the importance of negative design. Moreover, our approximation of the free energy of the misfolded state (Minning *et al.*, 2013) allow us to estimate the effect of each mutation on both unfolding and misfolding performing a number of computations that grows only linearly with the number of amino acids. As a consequence, these models can be applied along long phylogenetic histories.

The recently developed *CASS* tool (Grahnen and Liberles, 2012) can simulate protein sequence evolution accounting for selection to fold into a specific conformation. CASS is based on an all-atoms

representation of the protein and adopts an atomistic force fields, which can make it more accurate than our contact-matrix based method (although it is known that contact matrices allow to reconstruct all atoms coordinates with precision), but limits its treatment of the misfolded ensemble, which is important to avoid bias towards hydrophobic sequences that are often unrealistically favored by energy functions, although additional considerations in the latter might avoid this effect. An important characteristic of CASS, not included in our program, is that it allows selecting for structures that bind a target molecule, therefore allowing investigating an important aspect of protein function (which is, nevertheless, intimately related with structural stability (e.g., Lukatsky *et al.*, 2007)).

Another important feature of our framework, that is not present in the CASS approach, is that it allows modeling evolutionary mechanisms other than point mutations, such as recombination, which has been shown to be a key element in protein engineering (Carbone and Arnold, 2007) and could affect structural constraints (Archer *et al.*, 2008; Simon-Loriere *et al.*, 2009). For example, Xu *et al.* (2005) have shown that recombination may influence structural divergence. In addition, we can simulate molecular evolution in population genetics scenarios including demographics, population structure and migration, which allows to address a wide range of biological problems by investigating how these evolutionary scenarios influence the properties of the evolved sequences by altering the underlying genealogies and the properties of the substitution process, which in non-neutral fitness landscapes strongly depend on population size. But also the opposite, how including or ignoring structural considerations can affect population genetic inferences (i.e., inferences on recombination, demographics or migration).

We assessed the performance of the SCS models versus more traditional empirical replacement matrices. For all 10 protein families studied and tested combinations of temperature and configurational entropy parameters, the neutral SCS model always simulated sequences with amino acid frequencies closer to the observed ones than the best-fit empirical amino acid replacement model. On the other hand, the fitness SCS model may outperform the neutral model under optimal thermodynamic parameters, but it is very sensitive to the correct choice of parameters. Consequently, we recommend the use of the neutral SCS substitution model by default since it is more robust, while the fitness SCS model should be only used with the default thermodynamic parameters, in which case it may simulate more realistic proteins for some protein families. This difference stems from the fact that the non-neutral model optimizes protein stability when the population size is large, and it may bias the evolutionary process if the thermodynamic model is not reliable. On the other hand, we think that this dependence on parameters is reassuring, since it shows that the agreement between observed and simulated distributions that *ProteinEvolver* achieves is not a trivial result.

The benefits of using the SCS substitution models instead of the empirical substitution models were also observed by evaluating the adequacy between the simulated sequence and its best homology model with the DOPE energy. In particular, we found that the DOPE energy from proteins simulated by the SCS neutral model was always more negative (more stable three dimensional proteins) than proteins simulated under the empirical substitution model, although, not surprisingly, less negative than the energy of the experimentally observed sequence-structure pair. These findings were expected because the empirical models consider independent sites and therefore they are unable to account for physical interactions that promote stability (e.g., Pollock *et al.*, 2012; Rodrigue *et al.*, 2005). Nevertheless, they constitute a minimal test that indicates the consistency of SCS models and confirms the limitation of empirical substitution matrices.

The consideration of structural information should result in more sensitive and more accurate representations of molecular evolution than those based on sequence data alone (Wilke, 2012). Our structure-based models could help for a more realistic benchmarking of methods trying to take into account site-dependency induced by protein structure (e.g., Grahnen *et al.*, 2011; Nasrallah *et al.*, 2011) and the important influences of the unfolded and misfolded configurations on the protein stability (Bastolla and Demetrius, 2005; Mendez *et al.*, 2010; Zheng *et al.*, 2013). At the population level, the framework may help, for example: (i) to evaluate the range of proteins that one may expect to observe in different populations (where these populations can change their sizes with time and can exchange migration), (ii) to validate analytical frameworks (for example, methods for the inference of ARGs, ancestral protein reconstruction, recombination breakpoints and recombination rates, from proteins or protein-coding genes while accounting for structural constrains) or even (iii) to infer evolutionary parameters of interest and carry out model choice in an Approximate Bayesian Computation (ABC) approach (Beaumont et al., 2002); for example, estimate recombination rates or select among different demographic and migration models from protein data while accounting for structural information. At the molecular level, the framework may help, for example, to study the influence of recombination events on the structure-based stability of the resulting proteins or to perform structurally constrained substitution model choice by using ABC.

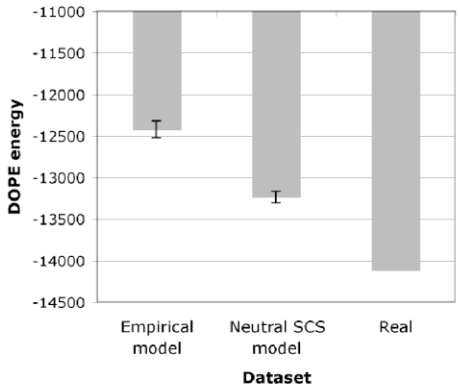


Fig. 3. DOPE energy computed in the simulated proteins under the empirical and the neutral SCS substitution models and in the native protein, for the protein family "Phototactive yellow proteins". Note that the DOPE energy is unnormalized with respect to the protein size and therefore scores from different proteins cannot be compared directly.

ACKNOWLEDGEMENTS

We want to thank David Abia for constructive comments and useful suggestions made during this study.

Funding: This work was supported by the Spanish Government with the "Juan de la Cierva" fellowship JCI-2011-10452 to MA

and grants BFU2011-24595 and BFU2012-40020 to UB. DP was financially supported by the European Research Council (ERC-2007-Stg 203161-PHYGENOM).

REFERENCES

Abascal, F., *et al.* (2005) ProtTest: selection of best-fit models of protein evolution, *Bioinformatics*, **21**, 2104-2105.

Anisimova, M. and Liberles, D.A. (2007) The quest for natural selection in the age of comparative genomics, *Heredity*, **99**, 567-579.

Anisimova, M., *et al.* (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites, *Genetics*, **164**, 1229-1236.

Archer, J., *et al.* (2008) Identifying the important HIV-1 recombination breakpoints, *PLoS Comput Biol*, **4**, e1000178.

Arenas, M. (2012) Simulation of Molecular Data under Diverse Evolutionary Scenarios, *PLoS Comput Biol*, **8**, e1002495.

Arenas, M. (2013) Computer programs and methodologies for the simulation of DNA sequence data with recombination, *Front Genet*, **4**, 9.

Arenas, M. and Posada, D. (2007) Recodon: coalescent simulation of coding DNA sequences with recombination, migration and demography, *BMC Bioinformatics*, **8**, 458.

Arenas, M. and Posada, D. (2010a) Coalescent simulation of intracodon recombination, *Genetics*, **184**, 429-437.

Arenas, M. and Posada, D. (2010b) The effect of recombination on the reconstruction of ancestral sequences, *Genetics*, **184**, 1133-1139.

Bastolla, U. and Demetrius, L. (2005) Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds, *Protein Eng Des Sel*, **18**, 405-415.

Bastolla, U., *et al.* (2003) Statistical properties of neutral evolution, *J. Mol. Evol.*, **57 Suppl 1**, S103-119.

Bastolla, U., *et al.* (2005a) Looking at structure, stability, and evolution of proteins through the principal eigenvector of contact matrices and hydrophobicity profiles, *Gene*, **347**, 219-230.

Bastolla, U., *et al.* (2005b) Principal eigenvector of contact matrices and hydrophobicity profiles in proteins, *Proteins*, **58**, 22-30.

Bastolla, U., *et al.* (2006) A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank, *BMC Evol. Biol.*, **6**, 43.

Bastolla, U., *et al.* (1999) Neutral evolution of model proteins: diffusion in sequence space and overdispersion, *J. Theor. Biol.*, **200**, 49-64.

Beaumont, M.A., *et al.* (2002) Approximate Bayesian computation in population genetics, *Genetics*, **162**, 2025-2035.

Berard, J. and Gueguen, L. (2012) Accurate estimation of substitution rates with neighbor-dependent models in a phylogenetic context, *Syst. Biol.*, **61**, 510-521.

Berezovsky, I.N., *et al.* (2007) Positive and negative design in stability and thermal adaptation of natural proteins, *PLoS Comput Biol*, **3**, e52.

Carbone, M.N. and Arnold, F.H. (2007) Engineering by homologous recombination: exploring sequence and function within a conserved fold, *Curr. Opin. Struct. Biol.*, **17**, 454-459.

Cruzan, M.B. (2001) Population size and fragmentation thresholds for the maintenance of genetic diversity in the herbaceous endemic Scutellaria montana (Lamiaceae), *Evolution*, **55**, 1569-1580.

Derrida, B. (1981) Random Energy Model: An exactly solvable model of disordered systems, *Phys Rev B*, **24**, 2613-2626.

Eswar, N., *et al.* (2006) Comparative protein structure modeling using Modeller, *Curr Protoc Bioinformatics*, **Chapter 5**, Unit 5 6.

Ewens, W.J. (1979) *Mathematical Population Genetics*. Springer-Verlag, Berlin.

Goldstein, R.A. (2011a) The evolution and evolutionary consequences of marginal thermostability in proteins, *Proteins*, **79**, 1396-1407.

Goldstein, R.A. (2011b) The evolution and evolutionary consequences of protein marginal stability, *Proteins*, **In press**.

Grahnen, J.A. and Liberles, D.A. (2012) CASS: Protein sequence simulation with explicit genotype-phenotype mapping, *Trends in Evolutionary Biology*, **4**, 1.

Grahnen, J.A., *et al.* (2011) Biophysical and structural considerations for protein sequence evolution, *BMC Evol. Biol.*, **11**, 361.

Griffiths, R.C. and Marjoram, P. (1997) An ancestral recombination graph. In Donelly, P. and Tavaré, S. (eds), *Progress in population genetics and human evolution*. Springer-Verlag, Berlin, 257-270.

Guerois, R., *et al.* (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations, *J. Mol. Biol.*, **320**, 369-387.

Hoban, S., *et al.* (2012) Computer simulations: tools for population and evolutionary genetics, *Nat. Rev. Genet.*, **13**, 110-122.

Hudson, R.R. (1983) Properties of a neutral allele model with intragenic recombination, *Theor. Popul. Biol.*, **23**, 183-201.

Hudson, R.R. (1998) Island models and the coalescent process, *Mol Ecol*, **7**, 413-418.

Javier Zea, D., *et al.* (2013) Protein conformational diversity correlates with evolutionary rate, *Mol. Biol. Evol.*, **30**, 1500-1503.

Juritz, E., *et al.* (2013) Protein conformational diversity modulates sequence divergence, *Mol. Biol. Evol.*, **30**, 79-87.

Kingman, J.F.C. (1982) The coalescent, *Stochastic Processes and their Applications*, **13**, 235-248.

Kosakovsky Pond, S.L., *et al.* (2008) Estimating selection pressures on HIV-1 using phylogenetic likelihood models, *Stat. Med.*, **27**, 4779 - 4789.

Krishna, M.M., *et al.* (2004) Protein misfolding: optional barriers, misfolded intermediates, and pathway heterogeneity, *J. Mol. Biol.*, **343**, 1095-1109.

Liberles, D.A., *et al.* (2012) The interface of protein structure, protein biophysics, and molecular evolution, *Protein Sci.*, **21**, 769-785.

Liò, P. and Goldman, N. (1998) Models of molecular evolution and phylogeny, *Genome Res.*, **8**, 1233-1244.

Lukatsky, D.B., *et al.* (2007) Structural similarity enhances interaction propensity of proteins, *J. Mol. Biol.*, **365**, 1596-1606.

Marti-Renom, M.A., *et al.* (2000) Comparative protein structure modeling of genes and genomes, *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291-325.

Mendez, R., *et al.* (2010) Mutation bias favors protein folding stability in the evolution of small populations, *PLoS Comput Biol*, **6**, e1000767.

Minning, J., *et al.* (2013) Detecting selection for negative design in proteins through an improved model of the misfolded state, *Proteins*.

Nasrallah, C.A., *et al.* (2011) Quantifying the impact of dependent evolution among sites in phylogenetic inference, *Syst. Biol.*, **60**, 60-73.

Neuhauser, C. and Tavaré, S. (2001) The coalescent, *Encyclopedia of Genetics*, **I**, 392-397.

Noivirt-Brik, O., *et al.* (2009) Trade-off between positive and negative design of protein stability: from lattice models to real proteins, *PLoS Comput Biol*, **5**, e1000592.

Nordborg, M. (2000) Coalescent Theory, *Review*.

Oostermeijer, J.G.B., *et al.* (1994) Offspring fitness in relation to population size and genetic variation in the rare perennial plant species Gentiana pneumonanthe (Gentianaceae), *Oecologia*, **97**, 289-296.

Parisi, G. and Echave, J. (2001) Structural constraints and emergence of sequence patterns in protein evolution, *Mol. Biol. Evol.*, **18**, 750-756.

Pollock, D.D., *et al.* (2012) Amino acid coevolution induces an evolutionary Stokes shift, *Proc Natl Acad Sci U S A*, **109**, E1352-1359.

Posada, D. (2001) The effect of branch length variation on the selection of models of molecular evolution, *J. Mol. Evol.*, **52**, 434-444.

Posada, D., *et al.* (2002) Recombination in evolutionary genomics, *Annu. Rev. Genet.*, **36**, 75-97.

Rastogi, S., *et al.* (2006) Evaluation of models for the evolution of protein sequences and functions under structural constraint, *Biophys. Chem.*, **124**, 134-144.

Rodrigue, N., *et al.* (2005) Site interdependence attributed to tertiary structure in amino acid sequence evolution, *Gene*, **347**, 207-217.

Sali, A. and Blundell, T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints, *J. Mol. Biol.*, **234**, 779-815.

Schierup, M.H. and Hein, J. (2000) Recombination and the molecular clock, *Mol. Biol. Evol.*, **17**, 1578-1579.

Sella, G. and Hirsh, A.E. (2005) The application of statistical physics to evolutionary biology, *Proc Natl Acad Sci U S A*, **102**, 9541-9546.

Shen, M.Y. and Sali, A. (2006) Statistical potential for assessment and prediction of protein structures, *Protein Sci.*, **15**, 2507-2524.

Simon-Loriere, E., *et al.* (2009) Molecular mechanisms of recombination restriction in the envelope gene of the human immunodeficiency virus, *PLoS Pathog*, **5**, e1000418.

Slatkin, M. (1996) Gene genealogies within mutant allelic classes, *Genetics*, **143**, 579-587.

Taverna, D.M. and Goldstein, R.A. (2002a) Why are proteins marginally stable?, *Proteins*, **46**, 105-109.

Taverna, D.M. and Goldstein, R.A. (2002b) Why are proteins so robust to site mutations?, *J. Mol. Biol.*, **315**, 479-484.

Vendruscolo, M., *et al.* (1997) Recovery of protein structure from contact maps, *Fold Des*, **2**, 295-306.

Wakeley, J. (2008) *Coalescent Theory: An Introduction*. Roberts and Company Publishers, Greenwood Village, Colorado.

Wilke, C.O. (2012) Bringing molecules back into molecular evolution, *PLoS Comput Biol*, **8**, e1002572.

Wiuf, C. and Posada, D. (2003) A coalescent model of recombination hotspots, *Genetics*, **164**, 407-417.

Xu, Y.O., *et al.* (2005) Divergence, recombination and retention of functionality during protein evolution, *Hum Genomics*, **2**, 158-167.

Yang, Z. (2006) *Computational Molecular Evolution*. Oxford University Press.

Zheng, W., *et al.* (2013) Frustration in the energy landscapes of multidomain protein misfolding, *Proc Natl Acad Sci U S A*, **110**, 1680-1685.

4.3.- Estudio del mimetismo molecular entre péptidos de *Chlamydia trachomatis* y péptidos humanos en la interacción con HLA-B27 y su importancia en la artritis reactiva

4.3.1.- Introducción y aportación del autor

El complejo mayor de histocompatibilidad (MHC) de clase I puede unir péptidos derivados de patógenos intracelulares y péptidos de proteínas propias, exponiéndolos en la superficie celular para su reconocimiento por los linfocitos T citotóxicos (CTLs). El alotipo o variante alélica HLA-B27 en humanos, de distribución mundial y diana del presente trabajo, muestra una fuerte asociación con un grupo de enfermedades reumatológicas denominadas espondiloartropatías. Una de ellas, la artritis reactiva, podría ser desencadenada por diversas bacterias (artritogénicas), siendo *Chlamydia. Trachomatis (C. Trachomatis),* un patógeno intracelular obligado*,* la más prominente. Cuando el complejo MHC, a consecuencia de una infección, presenta los péptidos bacterianos (antígenos) a los CTLs, se desencadena la respuesta inmune pudiendo generar memoria a largo plazo y estar relacionada con las artropatías. A pesar de que el papel patogénico de HLA-B27 en estas enfermedades aún no se ha identificado, uno de los mecanismos propuestos es el de la reactividad cruzada por parte de los CTLs, donde péptidos humanos y bacterianos al ser presentados por HLA-B27 adoptarían conformaciones similares. Este mecanismo de mimetismo molecular provocaría la respuesta autoinmune. Para evaluar esta hipótesis *in silico* nos planteamos estudiar a nivel atómico la similitud conformacional y la flexibilidad de los diferentes péptidos unidos a HLA-B*27:05 cuya presencia se ha identificado experimentalmente.

El trabajo que presentamos en el artículo (Alvarez-Navarro et al., 2013) recoge una variedad de péptidos de *C. trachomatis* que presentan homología de secuencia con proteínas humanas, entre los cuales 2 péptidos de la DNA primasa (DNAP), DNAP(211-221) y DNAP(211-223), presentan homología con un péptido derivado de la propia molécula HLA-B27 denominado B27(309-320) y se incluirán en el análisis bioinformático llevado a cabo por la autora de la tesis usando técnicas de HM, MD y *clustering* de las conformaciones obtenidas de los péptidos.

En primer lugar, dado que carecíamos de las estructuras resueltas experimentalmente con los péptidos unidos en su sitio activo, llevé a cabo el HM del complejo MHC-I, incluyendo la cadena de HLA-B*27:05 que es donde se localiza el sitio de unión de los péptidos, la β2-microglobulina y los respectivos péptidos. La construcción de los péptidos se realizó en base al

alineamiento múltiple de una colección de secuencias de péptidos cristalizados con HLA-B*27:05, manteniendo fijas las posiciones de anclaje y permitiendo variabilidad en la región central, más o menos expuesta al solvente según la longitud de sus secuencias. Los complejos obtenidos fueron refinados mediante MD en una caja de aguas incluyendo iones de neutralización. Tras ello realicé un análisis exhaustivo a nivel estructural y energético de cada uno de los complejos estudiados. Mediante el *clustering* de las estructuras de los péptidos recogidas a lo largo del tiempo obtuvimos una serie de estructuras representativas de las cuales comparamos sus conformaciones y sus distribuciones de carga. Observamos diversos grados de flexibilidad en los péptidos y cierta similitud en sus interacciones con la proteína HLA-B27 y en la parte expuesta de reconocimiento por el TCR o receptor de linfocitos T correspondiente. El péptido B27(309-320) se muestra como el más flexible de todos con 3 estructuras representativas principales que agrupan el 49,5%, 35,5% y el 11,2% de las conformaciones obtenidas con MD. Por otro lado, DNAP(211-223) y DNAP(211-221) presentan una mayor rigidez. El 99,7% de las conformaciones del péptido DNAP(211-223) observadas durante su dinámica molecular se agruparon en un solo *cluster*, mientras que las conformaciones de DNAP(211-221) se agruparon en 2 *clusters* representativos incluyendo el 14,7% y el 83,9% de las conformaciones observadas en la simulación. Hemos evidenciado que los péptidos de *C. trachomatis* presentan una menor flexibilidad que el péptido endógeno y que decrece aún más a medida que aumenta la longitud de su cadena a pesar de encontrarse más expuesto al solvente, siendo estabilizados mediante interacciones con las hélices que rodean el sitio activo.

Nuestros resultados sugieren la existencia de un mimetismo molecular entre los péptidos derivados de *C. trachomatis* y los péptidos de HLA-B27 debido tanto a su similitud en la distribución de cargas expuestas al TCR como a la alta flexibilidad de B27(309-320) en su región central para adquirir conformaciones cercanas a las de los péptidos de *C. trachomatis*, lo cual podría favorecer la reactividad cruzada de los CTLs frente a estos ligandos. Finalmente, esta reactividad cruzada tendría un papel muy importante en el desarrollo y/o cronicidad de la artritis reactiva.

Artículo 3

JBC Papers in Press. Published on July 18, 2013 as Manuscript M113.493247
The latest version is at http://www.jbc.org/cgi/doi/10.1074/jbc.M113.493247

Novel HLA-B27-restricted epitopes from *Chlamydia trachomatis* generated upon endogenous processing of bacterial proteins suggest a role of molecular mimicry in reactive arthritis*

Carlos Alvarez-Navarro[1], Juan J. Cragnolini[1,3], Helena G. Dos Santos[1], Eilon Barnea[2], Arie Admon[2], Antonio Morreale[1,4], and José A. López de Castro[1].

[1] From the Centro de Biología Molecular Severo Ochoa (Consejo Superior de Investigaciones Científicas and Universidad Autónoma), Madrid, Spain
[2]Faculty of Biology, Technion - Israel Institute of Technology, Haifa 32000, Israel.
[3]Current Address: Whitehead Institute for Biomedical Research, Cambridge, USA.
[4]Current address: Repsol, Technology Center, Móstoles, Madrid, Spain.

*Running title: *Chlamydial HLA-B27 ligands*

To whom correspondence should be addressed: Dr. José A. López de Castro. Centro de Biología Molecular Severo Ochoa, c/ Nicolás Cabrera N. 1, Universidad Autónoma, 28049 Madrid, Spain. Phone: 34-91-196 4554; Fax: 34-91-196 4420. Email: aldecastro@cbm.uam.es

Keywords: Antigen processing; Antigen presentation; Chlamydia; Molecular dynamics; MHC; Arthritis; HLA-B27; Pathogenesis; Autoimmunity.

Background: Reactive arthritis is an HLA-B27-associated disease triggered by *Chlamydia trachomatis*
Results: Three chlamydial peptides endogenously presented by HLA-B27 were identified. All were homologous to human-derived sequences and one showed conformational similarity to a self-derived HLA-B27 ligand.
Conclusion: Molecular mimicry between chlamydial and self-derived HLA-B27 ligands is not uncommon.
Significance: Molecular mimicry may contribute to the pathology of reactive arthritis.

ABSTRACT

Reactive arthritis (ReA) is an HLA-B27-associated spondyloarthropathy that is triggered by diverse bacteria, including *Chlamydia trachomatis*, a frequent intracellular parasite. HLA-B27-restricted T-cell responses are elicited against this bacterium in ReA patients, but their pathogenetic significance, autoimmune potential, and relevant epitopes are unknown. High resolution and sensitivity mass spectrometry was used to identify HLA-B27 ligands endogenously processed and presented by HLA-B27 from three chlamydial proteins for which T-cell epitopes were predicted. Fusion protein constructs of ClpC, (Na+)-translocating NADH-quinone reductase subunit A, and DNA primase were expressed in HLA-B27$^+$ cells and their HLA-B27-bound peptidomes were searched for endogenous bacterial ligands. A non-predicted peptide, distinct from the predicted T-cell epitope, was identified from ClpC. A peptide recognized by T cells *in vitro*, NQRA(330-338), was detected from the reductase subunit. This is the second HLA-B27-restricted T-cell epitope from *C. trachomatis* with relevance in ReA demonstrated to be processed and presented in live cells. A novel peptide from the DNA primase, DNAP(211-223), was also found. This was a larger variant of a known epitope and was highly homologous to a self-derived natural ligand of HLA-B27. All three bacterial peptides showed high homology with human sequences containing the binding motif of HLA-B27. Molecular dynamics simulations further showed a striking conformational similarity between DNAP(211-223) and its homologous and much more flexible human-derived HLA-B27 ligand. The results suggest that molecular mimicry between HLA-B27-restricted bacterial and self-derived epitopes is frequent and may play role in ReA.

MHC class I (MHC-I)[1] molecules present endogenous peptides derived from self-

proteins or intracellular pathogens at the cell surface for recognition by cytotoxic T lymphocytes (CTL). HLA-B27, an allotype that is present worldwide, shows one of the strongest associations between MHC-I and a human disease (1-3). This association concerns a group of inflammatory rheumatic diseases termed spondyloarthropathies, which include ankylosing spondylitis (AS), where this allele occurs in about 90% of patients, and reactive arthritis (ReA), where the prevalence of HLA-B27 is less well defined, but probably around 30-50% (4). This later disorder is triggered by various gram-negative bacteria (5). Although it is frequently a self-limited disease, ReA evolves sometimes towards AS, particularly among HLA-B27$^+$ individuals. In contrast to AS where HLA-B27 is probably a true pathogenetic factor, epidemiologic and other studies suggest that in ReA it may influence the severity of clinical manifestations, rather than being a truly causative allele (4;6;7).

Chlamydia trachomatis is a major agent in sexually transmitted infections (8). It is often asymptomatic, highly persistent and difficult to detect by conventional diagnostic tests. It is an obligate intracellular pathogen, which infects mucosal epithelial cells, vascular endothelial and other cells, such as monocytes and macrophages (9) and is one of the main pathogenetic agents in ReA.

C. trachomatis has developed multiple strategies to evade the immune system, including modulation of host cell apoptosis (10-14) and replication inside a specialized vacuole, called the inclusion, which limits its exposure to antibodies and to the antigen processing machinery (15). A third mechanism is associated with secretion of IFN-γ by immune cells. This cytokine inhibits bacterial growth through deprivation of the tryptophan pool, which leads to bacterial persistence under sub-inhibitory IFN-γ concentrations (16;17). Finally, *C. trachomatis* secretes a protease into the cytosol of the infected cell, the chlamydial protease-like activating factor, that degrades transcription factors for MHC, inhibiting the expression of MHC-I and -II at the cell surface shortly after infection (18-21). Despite this, both CD4+ and CD8+-mediated immune responses are activated upon infection (22).

The pathogenetic role of HLA-B27 in spondyloarthropathies remains ill-defined. Among the various proposed mechanisms (23), the *arthritogenic peptide* hypothesis (24) claims that a bacterial peptide presented by HLA-B27 would elicit a CTL response cross-reactive with a self-derived B27 ligand showing antigenic mimicry, thus breaking the self-tolerance and triggering an autoimmune attack (25). Although this mechanism does not satisfactorily explain AS pathogenesis, since the HLA-B27-associated spondyloarthopathy in transgenic rats does not require CD8$^+$ T cells (26), it may well play a role in exacerbating the pro-inflammatory nature of HLA-B27, particularly in ReA. Indeed, splenocytes from rats immunized with HLA-B27 and stimulated *in vitro* with *Chlamydi*a-treated cells from HLA-B27 transgenic rats resulted in the generation of *Chlamydia*-specific CD8$^+$ T-cells (27). Moreover, splenocytes from HLA-B27 transgenic rats immunized with HLA-B27 developed HLA-B27-directed autoreactivity upon exposure to *C. trachomatis in vitro* (28). The immunological relationship between *Chlamydia* and HLA-B27 revealed by these studies was suggestive of molecular mimicry between bacterial and self-derived HLA-B27-restricted epitopes. In spite of difficulties in substantiating molecular mimicry as a mechanism of autoimmunity (29), it played a key role in the pathogenesis of *Chlamydia*-induced autoimmune myocarditis in mice (30). Thus, there is a sound basis to search for HLA-B27-restricted chlamydial T-cell epitopes and their possible relationship to self-derived HLA-B27 ligands (31).

Predictive binding and proteasomal cleavage algorithms were used to localize putative chlamydial epitopes. The candidates were tested for recognition by specific CTL from transgenic mice or HLA-B27$^+$ ReA patients (32), or used for generating B27 tetramers to detect peptide-specific T cells (33). These studies identified some HLA-B27-restricted epitopes for which specific CTL could be found in *Chlamydia*-infected ReA patients. However, due to the intrinsic cross-reactivity of T cells (34), recognition of a synthetic peptide *in vitro* does not guarantee that this peptide is the actual immunogenic epitope *in vivo*.

The direct biochemical identification of endogenous chlamydial T-cell epitopes from infected cells has been accomplished only in the mouse system (35;36). It is hardly feasible in humans, due to the very low amounts of bacterial epitopes on infected cells, the difficulties associated to working with large amounts of *Chlamydia*-infected human cells and, specially, the downregulation of MHC-I

expression and induction of apoptosis by *C. trachomatis* (19;37). Thus, we developed an alternative strategy involving the stable expression of chlamydial fusion proteins on HLA-B27^{+} human cells. Endogenously processed chlamydial peptides, including a predicted T-cell epitope, were identified by comparing the HLA-B27-bound peptidomes from transfected and untransfected cells. These studies (38;39) were based on comparative MALDI-TOF MS and concerned 3 chlamydial proteins containing sequences highly homologous to known human-derived HLA-B27 ligands or from which synthetic peptides were recognized by CTL from ReA patients: DNA primase (DNAP) (CT794), Na^{+}-translocating NADH-quinone reductase subunit A (NQRA) (CT634) and pyrroloquinoline-quinone synthase-like protein (PqqC) (CT610).

In two different studies, based on a predictive search for HLA-B27-restricted chlamydial ligands in ReA patients (32;33), a sequence from ClpC protein, spanning residues 7-15, was recognized as a synthetic peptide by CD8+ T cells from multiple individuals, suggesting that this epitope could be immunodominant. Here we used MS techniques of high sensitivity and accuracy to investigate the endogenous processing and presentation of this and other HLA-B27-restricted peptides from ClpC and other chlamydial proteins. Molecular dynamics simulations were also carried out to analyze the relationship between chlamydial and homologous human-derived B27 ligands at the conformational level.

EXPERIMENTAL PROCEDURES

ClpC Gene Constructs-Enhanced GFP (EGFP)-ClpC fusion proteins were generated by fusing the cDNA of the *clpC* gene (CT286) of *C. trachomatis* serovar L2 (Advanced Biotechnologies, Columbia, MD) or truncated forms of it, in frame to the 3'-end of the EGFP gene. Full-length cDNA of ClpC was amplified by PCR using the following primers: 5'-CTCTCTCTAGATCTATGTTTGAGAAGTTT ACCAATCG and 3'-CTCTCTCTGTCGACCTATGATTCATCAGC TGTAATAG. The PCR products were cloned into the pEGFP-C1 vector (BD Biosciences Clontech) using 5' *BglII* and 3' *SalI* restriction sites. Two constructs were made based on the EGFP-CT286 plasmid sequence and the internal restriction sites *BglII* at 5' and *ApaI*, and *EcoRI* at 3', respectively.

Cell culture and transfections-Stable transfectants were generated as previously described (38). Briefly, The EGFP-ClpC constructs were co-transfected by electroporation in C1R-B*27:05 cells (40), with the RSV5 vector (41) containing the hygromycin resistance gene. The transfected cells were selected with 250µg/mL hygromycin (Invitrogen). All cell lines were cultured in RPMI 1640 medium, supplemented with 10% FBS, 200mM L-Gln, 25mM HEPES, streptomycin and penicillin.

Flow cytometry-The C1R transfectants were analyzed by measuring their EGFP-associated fluorescence. Briefly, $1x10^{6}$ cells were washed twice with 200µL of PBS and centrifuged at 1500 rpm for 5 min. The detection was carried out in a flow cytometer FACSCalibur (Beckton Dickinson). All data were acquired using CellQuestTM Pro v4.0.2 software (BD Bioscience) and analyzed using FlowJo v7.5 (Tree Star, Inc.).

Immunoprecipitation and Western blot-About, $2x10^{6}$ cells were lysed in 0.5% Igepal CA-630 (Sigma), 5 mM $MgCl_2$, 50 mM Tris-HCl, pH 7.4, containing protease inhibitors (Complete Mini, Roche Applied Science) for 30 min. After centrifugation, the lysate supernatants were precleared with anti-rabbit IgG beads (TrueBlot, eBioscience, San Diego, CA) and immunoprecipitated for 3 hr with the rabbit anti-GFP polyclonal antibody (A6455) (Invitrogen) coupled to anti-rabbit IgG beads, at 4 °C and continuous shaking. Immunoprecipitates were washed three times, denatured for 5 min in sample buffer, subjected to 10% SDS-PAGE and transferred overnight to a nitrocellulose membrane (Amersham Hybond-ECL, GE Healthcare, UK) at 20V and 4°C. The immunodetection was carried out using the A6455 antibody and horseradish peroxidase-conjugated anti-Rabbit IgG (TrueBlot, eBioscience, San Diego, CA) at 1:1000 and 1:5000 dilutions, respectively. Antibodies were diluted in blocking buffer containing 5% non-fat dry milk; 0.1% Tween 20; PBS pH 7.4. The immunoblots were developed using the ECL immunodetection system (Amersham Bioscience).

Isolation of HLA-B27-bound peptides-B*27:05-bound peptides were isolated from about $1x10^{10}$ or for some analyses, $1x10^{9}$ C1R-B*27:05 cells, as previously described (42). Briefly, cells were lysed in the presence of a cocktail of protease inhibitors (Complete, Roche

Applied Science). The soluble fraction was subjected to affinity chromatography using the W6/32 mAb (IgG2a; specific for a monomorphic HLA class I determinant) (43). HLA-B27-bound peptides were eluted with 0.1% aqueous TFA at room temperature, filtered through Centricon 3 devices (Amicon, Beverly, MA), concentrated and either used as a peptide pool, or subjected to reverse phase HPLC fractionation at a flow rate of 100 µl/min, as previously described (44). Fractions of 50 µl were collected and stored at -20ºC until use.

Synthetic Peptides-These were obtained using standard N-(9 fluorenyl)methoxycarbonyl chemistry and purified by HPLC. The correct m.w. of purified peptides was verified by MALDI-TOF MS.

MALDI-TOF MS-HPLC fractions were analyzed using a MALDI-TOF/TOF mass spectrometer (4800 Proteomics Analyzer, Applied Biosystems, Foster City, CA) as previously described (38) and processed using the Data Explorer software version 4.9 (Applied Biosystems).

Electrospray-LTQ-Orbitrap MS/MS-Peptide mixtures were desalted and concentrated with Micro-Tip reverse-phase columns and analyzed by µLC-MS/MS using an Orbitrap XL mass spectrometer (Thermo Fisher, San Jose, CA) fitted with a capillary HPLC (Eksigent, Dublin, CA) as previously described (45), with minor modifications. Briefly, the peptides were eluted at flow rates of 0.25 µl/min, with linear gradients of 7–40% acetonitrile in 0.1% formic acid, during 90 min, followed by 17 min at 95% acetonitrile in 0.1% formic acid. In some cases, the same gradient was used during 214 min, with a final isocratic elution for 29 min. The spectra were collected in the orbitrap mass analyzer using full ion scan mode over the mass-to-charge (m/z) range 400–2000, which was set to 60000 resolutions. The most intense 7 masses from each full mass spectrum, with single, double and triple charge states, were selected for fragmentation by collision-induced disintegration in the linear ion-trap.

Electrospray-LTQ-Velos MS/MS-Particular peptides were searched in 10 µl of individual HPLC fractions by MS/MS in a dual mode, using Selected Multiple Ion Monitoring (SMIM) and dynamic exclusion mode in an LTQ-Velos instrument. Briefly, each particular fraction was dried down and resuspended in 9 uL of 0.1% formic acid and analyzed in an Agilent 1100 system coupled a linear ion trap LTQ-Velos mass spectrometer (Thermo Fisher Scientific, Waltham, MA, USA). The peptides were separated by reverse phase chromatography using a 0.18×150 mm Bio-Basic C18 RP column (Thermo Fisher Scientific) and eluted using an 80-min gradient from 5 to 40% solvent B (Solvent A: 0.1% formic acid in water, solvent B 0.1% formic acid, 80% acetonitrile in water) at 1.8 µl/min. Peptides were detected in SMIM mode, at single, double and triple charged states. In parallel to the SMIM mode, a full ion scan over the *m/z* range 400-2000 (1 µscans) was also performed, followed by data dependent MS/MS scans, using an isolation width of 2 *m/z* units, normalized collision energy of 35%, and dynamic exclusion was applied during 30 seconds. Alternatively, 10 µl aliquots of various consecutive HPLC fractions were pooled together, and analyzed in the same way. The synthetic peptides were detected using only the SMIM mode as above, except that a 35-min elution gradient was used.

Database searches-The Mascot server 2.2 (Matrix Science Inc. Boston, USA) (46) was used as the main search engine. The search parameters were 0.5 Da mass tolerance for both precursor and fragment ions for MS/MS spectra from LTQ-Velos, and 0.01 Da and 0.5 Da for precursor and fragment ions, respectively, for data from LTQ-Orbitrap. Met oxidation, Asn and Gln deamidation were selected as variable modifications. A small sequence database consisting of the chlamydial ClpC (Swiss-Prot accession B0B7K2), DNAP (B0B920), NQRA (O84639) sequences, as well as HLA-B27 (P03989), HLA-B35 (P30685), HLA-C04 (P30504) and EGFP (Genbank accession AAB02576.1) was used for the specific search of chlamydial peptides. In addition, all raw files were run against the human subset of the Uniprot database (Release 57.6, 07/2009, with 20331 entries), using the same parameters described above. Those sequences showing the highest scores in these preliminary searches were analyzed manually and validated by comparison with the experimental MS/MS spectrum of the corresponding synthetic peptide.

The search for homology between chlamydial peptides and human proteins was carried out using the UniProtKB/Swiss-Prot database (release 07/2012, with 20231 entries) and the BLASTP 2.2.26+ software (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Proteasome cleavage predictions-Proteasome/immunoproteasome cleavage was

predicted with previously described algorithms (47) available at http://imed.med.ucm.es/Tools/pcps/index.html.

Homology Modeling (HM)-Three-dimensional models for the complexes between B*27:05/β2m and DNAP(211-221), DNAP(211-223) or B27(309-320) were built by HM. A total of 23 X-ray structures of HLA-B27/peptide complexes were aligned using the MAFFT software (48). Since all the X-ray complexes contained bound 9-mers, the alignments of these peptides with the longer ones in our study was done by introducing gaps at internal peptide positions. The 4 N-terminal and 2 C-terminal positions on each peptide were constrained, while certain flexibility was allowed for their central parts. B*27:05 in complex with the pVIPR(400-408) peptide in its canonical conformation (PDB code 1OGT) (49) was finally selected as template, due to its high resolution (1.47 Å), and the alignment was subjected to HM using the MODELLER program

Set up of the systems and molecular dynamics (MD) simulations-For each HLA-B27/peptide complex the set up entailed the following steps: a) adding missing heavy and hydrogen atoms (50) to assign atom types and charges according to AMBER ff10 force field (51) and to determine the protonation state of ionizable residues at pH=7; b) employing the tleap module from the AmberTools package (52) to immerse each system within a 10 Å box of TIP3P (53) explicit water molecules and to add Na^+ counter ions; c) the positions of water molecules and ions were energy minimized using the conjugated gradient method for 3000 steps while the atomic coordinates in the complexes were kept constrained, and then equilibrated at 298 K during 10 ps maintaining the constraints; d) the constraints were transformed into progressively lower restraints and the whole complexes, including the water molecules and the ions, were energy minimized as above.

MD simulations were carried out starting from the energy minimized structures. All calculations were performed with the NAMD 2.8 program (54) using constant temperature (298 K) and pressure (1 atm). Short and long-range forces were calculated every one and two time steps, respectively (a time step=2.0 fs), constraining the covalent bonds involving hydrogen atoms to their equilibrium values. Long-range electrostatic interactions were accounted for using the particle mesh Ewald approach (55) . The systems were heated up to 298 K and then equilibrated at this temperature during 200 ps. The equilibration was performed under harmonic restraint conditions on all the heavy atoms that were gradually reduced until they were almost removed. Finally, these equilibrated structures were further simulated during additional 50 ps with a minimal restraint. These were the starting points for a 30 ns MD production period during which the system coordinates were collected every 2 ps for further analysis.

Analysis of MD trajectories-The stability of a given complex was evaluated by calculating the root-mean-square deviation (RMSD) of the Cα atoms along the trajectories using as reference their starting structures. Additionally, the root-mean-square fluctuation (RMSF) of each residue was calculated using as reference their average value once each snapshot had been fitted to their initial structure. Further analysis was carried out by clustering the sampled conformational space during the trajectories production period (last 10 ns), using the ptraj module from the AmberTools package, the snapshots sampled as described above and the average-linkage algorithm based on the peptide backbone atoms. Adaptive Poisson-Boltzmann Solver (56;57) was used to perform the Poisson-Boltzmann electrostatic calculations for the most representative structures in each cluster. Dielectric constants were set to 4 and 80 for protein and solvent, respectively. Other parameters were set as default. The free energy of binding between each peptide and the B*27:05 molecule was calculated by the MM-ISMSA approach (58). We also calculated the pairwise decomposition of the free energy of binding following the scheme developed in MM-ISMSA to determine the main residues responsible for the interaction of the peptides with B*27:05. Mean and standard deviation for the free energy of binding was calculated for the MD trajectories fit to a normal frequency distribution using R (59). Contacts between residues were analyzed following the MM-ISMSA methodology.

RESULTS

Expression of Chlamydial ClpC fusion proteins-ClpC is an ATP-dependent protein-unfolding subunit of the bacterial ClpCP protease complex (60;61). In *C. trachomatis* it has 854 amino acid residues and binds ATP

through two nucleotide-binding domains, AAA+ (**Fig. 1A**). EGFP-ClpC fusion proteins were expressed in C1R-B*27:05 cells in order to detect endogenously processed HLA-B27 ligands from this protein, including a predicted T-cell epitope, ClpC(7-15). Our initial attempts to express the whole ClpC protein using full-length cDNA failed to generate stable C1R transfectants. To avoid functional interference of the ClpC protein in human cells, two fusion protein constructs, ClpC(1-570) and ClpC(1-512), with partial or total deletions of the C-terminal AAA+ domain, were made in which residues 1-570 or 1-512, respectively, were fused at the C-terminal end of EGFP **(Fig. 1A).** Stable transfectants in C1R-B*27:05 cells were obtained for both constructs, whose expression levels and correct size were determined by flow cytometry (**Fig. 1B**) and Western blot (**Fig. 1C**), respectively. The ClpC(1-512) transfectant in C1R-B*27:05 was used for further experiments, due to its higher expression compared to ClpC(1-570).

*One ClpC-derived ligand distinct from the predicted T-cell epitope is endogenously presented by HLA-B*27:05 on C1R cells*-A first approach to search for endogenously processed ClpC-derived HLA-B27 ligands was the comparative analysis of HLA-B27-bound peptides from untransfected C1R-B*27:05 cells and the ClpC(1-512) transfectant, based on identity of chromatographic retention time (RT) and m.w., through systematic comparison of the MALDI-TOF MS spectra from correlated HPLC fractions. Although this strategy was successful in previous studies with other fusion proteins (38;39), it failed to identify any ClpC-derived peptides. Thus, two further approaches were undertaken (**Fig. 1D**). The first one involved high-throughput sequencing, using LTQ-Orbitrap MS/MS, performed on the unfractionated B27-bound peptide pool from ClpC(1-512)-transfected C1R-B*27:05 cells. The second one involved a targeted search for specific candidates in the fractionated B27-bound peptide pool performed on HPLC fractions at the RT±3 min of each of the corresponding synthetic peptides. The relevant HPLC fractions, either individually or pooled together, were subjected to MS/MS fragmentation of all ions corresponding to the *m/z* ratios of the candidate peptide, using a LTQ-Velos mass spectrometer.

The MS/MS spectra from the unfractionated B27 peptidome from the ClpC(1-512) transfectant obtained in the LTQ-orbitrap were searched against a small database including ClpC and a few other chlamydial proteins. Two putatively significant matches with sequences containing the canonic B27 binding motif R2 from ClpC were obtained. Manual inspection of the corresponding MS/MS spectra showed a good match with the theoretical fragmentation of only one of these sequences: SRLDPVIGR, spanning ClpC residues 203-211 (**Fig. 2A**). A search against the human proteome database did not show a match of this MS/MS spectrum with any human peptide. SRLDPVIGR did not match any human sequence upon Blast analysis, confirming the bacterial origin of this peptide.

We next determined whether this peptide was just overlooked in our previous MALDI-TOF comparison or was hidden by a co-eluting human HLA-B27 ligand. For this purpose, the RT of the synthetic peptide in the same chromatographic conditions was determined and the fractions corresponding to its RT±3 min were fragmented in a LTQ-Velos mass spectrometer. Parental ions with *m/z* 506.80 and 338.20, compatible with the $[M+2H]^{2+}$ and $[M+3H]^{3+}$ forms of the chlamydial SRLDPVIGR peptide, respectively, were detected in Fr.142. The MS/MS spectrum of the former ion showed virtual identity with those from the LTQ-Orbitrap and the synthetic peptide (**Fig. 2A**). This assignment was further confirmed by the identity of the MS/MS spectrum of the ion with m/z 338.20 with that of the $[M+3H]^{3+}$ ion of the synthetic SRLDPVIGR (**Fig. 2B**). Comparative MALDI-TOF analysis of Fr.142 and adjacent ones confirmed the presence of a co-eluting self-derived B*27:05 ligand, as revealed by an ion peak with *m/z* 1012.53, identical to the $[M+H]^{+}$ of SRLDPVIGR, in cells lacking the chlamydial fusion protein (data not shown). This explains our failure to detect this bacterial peptide by MALDI-TOF.

Since the Orbitrap-based sequencing described in the previous paragraph failed to detect the predicted T-cell epitope ClpC(7-15), NRAKQVIKL, an alternative approach was used for the specific search of this and the related peptide ClpC(7-17), NRAKQVIKLAK, which also has the B*27:05 binding motif, in the HPLC-fractionated HLA-B27-bound peptide pool from the ClpC(1-512) transfectant. Both peptides were synthesized and used for a targeted search (**Fig. 1D**), monitoring the *m/z* ratios corresponding to $[M+2H]^{2+}$ and $[M+3H]^{3+}$ ions of both peptides. These analyses failed to

show any reliable fragmentation compatible with ClpC(7-15) or ClpC(7-17).

Novel chlamydial peptides from other proteins processed and presented by HLA-B27 in live cells-Several chlamydial peptides endogenously processed and presented by HLA-B27 were identified in previous studies from our laboratory (38;39) by comparative MALDI-TOF MS of HPLC-fractionated B27-bound peptide pools from C1R-B*27:05 transfectants expressing chlamydial NQRA, PqqC or DNAP fusion protein constructs (**Table I**). Due to the limitations of this approach, revealed by our results on ClpC, a search for novel peptides from NQRA and DNAP was undertaken, using more sensitive MS techniques.

NQRA. The NQRA(330-338) peptide, MRDHTITLL, was recognized *in vitro*, as a synthetic peptide, by $CD8^+$ T cells from a ReA patient (32), but was not found in C1R-B*27:05 cells expressing the EGFP-NQRA(1-465) fusion protein in a MALDI-TOF-based study (39). Thus, the most intense ions in the full MS spectrum of the pooled fractions corresponding to the RT±3 min of the synthetic peptide in the fractionated HLA-B27-bound peptide pool from the EGFP-NQRA(1-465) transfectant, were subjected to MS/MS fragmentation. The MS/MS spectrum of one of the main ion peaks in the full MS scan, with *m/z* 558.33, was compatible with the $[M+2H]^{2+}$ species of the oxidized form of MRDHTITLL. Its correct assignment was confirmed by comparison with the MS/MS spectrum of the corresponding synthetic peptide in its oxidized form (**Fig. 3**). This result demonstrates the endogenous processing and presentation by HLA-B27 of the predicted chlamydial epitope NQRA(330-338) in NQRA transfectant cells. This is the second HLA-B27-restricted T-cell epitope with demonstrated relevance in *Chlamydia*-infected ReA patients that is shown to be generated in live cells.

DNAP. The unfractionated HLA-B27-bound peptide pool from C1R-B*27:05 transfected with the EGFP-DNAP-(90-450) fusion protein (38) was subjected to MS/MS analysis in a LTQ-Orbitrap mass spectrometer and searched against a small database including the chlamydial DNAP fusion protein sequence. A parental ion of *m/z* 508.62, compatible with DNAP(211-223): RRFKEGGRGGKYI was identified (**Fig. 4A**).This peptide was two residues longer than one previously found from this protein: DNAP(211-221) (**Table I**). Both sequences show high homology with a natural ligand of HLA-B27, arising from the endogenous processing of the HLA-B27 heavy chain, B27(309-320): RRKSSGGKGGSY (62). To confirm the tentative assignment from the Orbitrap analysis, a targeted search for this peptide (**Fig. 1D**) was carried out in the HPLC-fractionated B27-bound peptide pool from the DNAP transfectant, focusing on the *m/z* values corresponding to the $[M+H]^+$, $[M+2H]^{2+}$, and $[M+3H]^{3+}$ forms of DNAP(211-223). The analysis revealed the presence of this peptide as the charge variants $[M+3H]^{3+}$ (*m/z* 508.62) (**Fig 4A**) and $[M+2H]^{2+}$ (*m/z* 762.43) (**Fig. 4B**), whose identity was confirmed by comparison with the MS/MS spectra of the synthetic peptide.

High homology between the ClpC and NQRA-derived HLA-B27 ligands and human sequences-To explore the possible molecular mimicry between the B27-restricted peptides from *C. trachomatis* found in this study and putative self-derived HLA-B27 ligands, we looked for human sequences showing high homology to ClpC(203-211) and NQRA(330-338). The search was performed against the human proteome, looking for sequences containing >50% amino acid identity with the bacterial peptides and the main binding motif of HLA-B27 ligands, R2. Only human sequences with residues present among known HLA-B27 ligands (63;64) with a frequency >1% at the anchor P1, P3 and PΩ positions were considered. Multiple human sequences homologous to the ClpC- and NQRA-derived peptides were found (**Table II**). Most of the sequences showed predictive scores compatible with proteasome/immunoproteasome cleavage at their C-terminal residue (>0.5).

MD simulation of chlamydial DNAP and homologous human-derived HLA-B27 ligands-To explore the similarity of DNAP(211-221) and DNAP(211-223) with B27(309-320) at the three-dimensional level, comparative MD simulation of their interaction in complex with B*27:05 was carried out. The initial, energy minimized, three-dimensional structures of the complexes involving the 3 peptides, all built by HM, and pVIPR(400-408) in its canonical conformation, were subjected to MD simulations for 30 ns. After this time, the stability of the trajectories was analyzed. Both the mean Cα RMSD and the mean RMSF for the B*27:05 heavy chain and β2m were similar among the 3 complexes (**Fig. 5A-B).** In contrast, the mean RMSD and RMSF values for the peptides were more variable, spreading from 0.58 to 2.25 Å

and from about 2.0 to 2.4 Å, respectively, in the different complexes (**Fig. 6A-B**). Large RMSF values (above 3.0 Å) were observed for certain residues (**Fig. 6B**), such as R8 in DNAP(211-221) and G6, G7 and K8 in B27(309-320). The very low RMSD fluctuation of DNAP(211-223) after the first 5-10 ns of MD simulation and the smaller RMSF values, relative to DNAP(211-221) and B27(309-320), suggest a less flexible structure of the former peptide.

Clustering analysis reveals distict peptide flexibility and conformations-A total of 5000 structures sampled during the last 10 ns of the MD simulation were subdivided in up to 5 clusters on the basis of similarity (RSMD) in the peptide backbone. Two predominant clusters were found for DNAP(211-221), 1 for DNAP(211-223), 3 for B27(309-320) and 1 for the X-ray template (**Table III**). The distinct flexibility of the three peptides revealed by this analysis was further apparent upon considering the intra-cluster RMSD variability. This was calculated as the distance to the centroid, which is the average distance of all members of a cluster to its geometrical center. This parameter reflects the dispersion of data inside a given cluster. Smaller (0.43), intermediate (0.54) and larger values (0.7) were found for the major clusters of DNAP (211-223), DNAP(211-221) and B27(309-320), respectively. These results indicate that, in complex with B*27:05, B27(309-320) is highly flexible, DNAP(211-221) has less flexibility, and DNAP(211-223) is significantly rigid. The overall structure of the peptide binding site showed no significant differences among the various complexes. A set of 100 unclustered structures homogenously sampled at 100 ps intervals in each modeled complex from the last 10 ns of the trajectories is shown in **Fig. 6C**.

Representative structures (reps) from each of the main clusters observed in B27(309-320), DNAP(211-221), and DNAP(211-223) (**Table III**) illustrate the three-dimensional configuration preferences of the peptides in their bound states **(Fig. 6D).** For B27(309-320), rep1 and rep2 showed similar conformations, and small differences in their molecular surface, but rep4 was significantly different. For DNAP(211-221), the representative conformers of its two main clusters were very similar, and different from those of B27(309-320). In contrast, the only major cluster in DNAP(211-223) showed a striking similarity to B27(309-320), looking like an intermediate form of rep2 and rep4 of this peptide. DNAP(211-223) also showed a surface charge distribution with similarities to both rep2 and rep4 of B27(309-320) **(Fig. 6E).**

Binding energy- MM-ISMSA was used to estimate the total free energy of binding of the peptides in the binding groove of B*27:05 and the contribution of each peptide residue to the total free energy of binding. The N- and C-terminal residues each contributed ~20 kcal/mol to the total binding of each peptide. Residue 2 showed the highest contribution, ~25 kcal/mol, while the central regions of the peptides showed greater variation and a smaller contribution (**Fig. 5C**). These results are in full agreement with the known canonical interactions governing binding of MHC-I ligands.

DISCUSSION

Two issues were addressed in this study. First, the endogenous processing and presentation of predicted T-cell epitopes, recognized as synthetic peptides by CTL from *Chlamydia*-infected ReA patients. Second, the structural similarity between chlamydial and human-derived HLA-B27 ligands. Our approach was the direct identification of endogenously processed chlamydial peptides using high sensitivity and accuracy MS. Although, ideally, this search should be performed on *Chlamydia*-infected cells, this approach is virtually unfeasible in humans, due to induction of MHC-I downregulation and apoptosis (38). Some chlamydial proteins are injected into the cytosol through the type III secretion system (65-68). However, many others reach cytosolic cross-presentation pathways (69;70) after uptake of bacterial debris from infected cells undergoing apoptosis, and are subjected to proteasomal degradation, similarly as endogenous proteins. Thus, the endogenous processing of chlamydial fusion proteins is likely to mimic that in infected cells to a large degree, as confirmed by the direct identification of chlamydial T-cell epitopes using fusion proteins in this and a previous study (39). Yet, proteasome-independent pathways might also generate chlamydial MHC-I ligands after transfer of bacterial components following the fusion of inclusion-derived vesicles with the ER (71), and perhaps also through non-cytosolic cross-presentation pathways. Thus, some chlamydial antigens may not be revealed with our approach.

Although studies based on MALDI-TOF MS allowed us to identify several HLA-B27 ligands from *C. trachomatis*, the limitations of

this approach granted a more in-depth search using electrospray-based MS techniques, to look for novel chlamydial epitopes. In spite of the technical improvements, the direct identification of immunologically relevant bacterial peptides by biochemical methods is less sensitive than CTL, since these can recognize minute antigen amounts, down to a few copies, at the cell surface (72). Although the relatively high expression of bacterial fusion proteins results in the generation of many more copies of chlamydial peptides than on infected cells, partially compensating for the lower sensitivity of biochemical analyses, the different thresholds relative to CTL recognition must always be kept in mind.

Our study focused on three chlamydial proteins. For two of them, ClpC and NQRA, HLA-B27-restricted T-cell epitopes had been predicted (32;33). For the third one, DNAP, an endogenous peptide, DNAP(211-221), with high homology to a natural human-derived B27 ligand, was previously reported (38). Both the transcriptional profile (73) and the proteomic characterization of the *Chlamydia* life cycle (74) indicate that ClpC is expressed in the infectious elementary body and, at higher level, in the replicative but non-infectious reticulate body, and is upregulated by IFN-γ (75). The presence of ClpC in both developmental stages and its upregulation in an inflammatory context is compatible with the possibility that HLA-B27-restricted T cells, directed against epitopes from this protein, may be relevant in controlling both the bacterial infection and the development of ReA. Detection of NQRA in the elementary body, but not in the reticulate body, is likewise compatible with the possibility that peptides from this protein may trigger B27-restricted T-cell responses at early stages of the infection. The finding of HLA-B27-restricted T-cells against peptides from these proteins in ReA patients (32;33) is consistent with both their expression patterns and possible pathological relevance.

T-cell epitope assignments based on predictive algorithms have limitations that preclude a reliable identification of relevant antigens without their direct detection *in vivo*. These limitations are clear in the previous failure to predict some chlamydial B27 ligands that are endogenously processed and presented in live cells, including ClpC(203-211) identified in this study. Moreover, since monoclonal T cells can recognize many distinct peptides (34), T-cell recognition of a synthetic peptide *in vitro* does not necessarily identify the natural epitope. Conversely, the identification of chlamydial peptides processed and presented by HLA-B27 in live cells does not indicate their immunological relevance in the absence of their positive identification by T cells.

In spite of their limitations, prediction algorithms are useful for detecting epitopes generated *in vivo,* since they help in focusing MS-based search strategies towards specific peptides in complex pools, as demonstrated by our previous identification of an endogenous HLA-B27-restricted chlamydial T-cell epitope (39). Another predicted epitope, from NQRA, was found in the present study. Thus, NQRA(330-338) is the second known chlamydial T-cell antigen processed and presented in live cells by HLA-B27 and recognized by specific CTL from ReA patients. This demonstrates the similarity of epitope processing between fusion proteins and infected cells.

Our failure to detect the predicted T-cell epitope ClpC(7-15), in spite of its intensive search with highly sensitive techniques, must be interpreted with caution. We cannot rule out that this peptide might be present in our cell lines in very low amounts that challenge detection by MS but are still sufficient for T-cell recognition. With this possibility in mind, our results suggest that this peptide may be produced with low efficiency, if at all, *in vivo*.

C. trachomatis is a large organism and is potentially the source of many HLA-B27-restricted ligands. The use of fusion proteins necessarily limits our analysis to a few epitopes. Yet, the endogenous generation of HLA-B27 ligands from each bacterial protein tested, suggests that HLA-B27-restricted T-cell responses in ReA patients may be directed against multiple chlamydial antigens. That all of the reported peptides showed significant homology with human sequences suggests that autoimmune cross-reaction of *Chlamydia*-specific T-cells with self-derived HLA-B27 epitopes through molecular mimicry might not be uncommon.

The chlamydial DNAP shows a particularly interesting example of molecular mimicry between bacterial and self-derived HLA-B27 ligands. HLA-B27 presents an 11-mer from this protein, DNAP(211-221), with high homology to the human-derived HLA-B27 ligand B27(309-320), which is one residue

longer than the chlamydial peptide (38;62). The finding now of the C-terminally extended variant DNAP(211-223), whose proteasomal generation was predicted in a previous study (62), increased the probability of molecular mimicry between peptides from DNAP and the human-derived ligand. MD simulations suggest that DNAP(211-221) and DNAP(211-223) adopt distinct conformations. Both peptides showed limited flexibility and a peptide-specific predominant conformation. In contrast, B27(309-320) was significantly more flexible. This is in agreement with X-ray data showing a single defined conformation of DNAP(211-221) and a diffuse electron density corresponding to the central region of B27(309-320) in complex with B*27:05 (B. Loll, B. Uchanska-Ziegler and A. Ziegler, unpublished observations, cited with permission). The limited flexibility of the two chlamydial peptides, specially DNAP(211-223), observed in our MD simulations was apparently determined by intra-peptide hydrogen bonds established within their central regions, which are more frequent among long peptides, and by peptide-specific interactions of their central regions with HLA-B27 residues.

The higher flexibility of the human-derived peptide is likely to provide a wider spectrum of antigenically distinct conformations. The striking similarity of the conformation and surface charge distribution of DNAP(211-223) with some of the main conformational clusters of B27(309-320) could favor T-cell cross-reaction between both peptides. A peptide bound in a flexible and variable conformation in its middle part may be amenable to recognition by more T-cell clones, with preference for single conformations, than a peptide bound with lower flexibility. For instance, T-cell mediated self-reactivity has been related to peptide antigens bound to HLA-B27 in dual conformation (76;77). The antigenic similarity between the DNAP-derived peptides and the homologous self-derived B27 ligand must be confirmed in functional assays with peptide-specific T cells. Although we recognize the importance of functional studies in this context, we were unable to perform them, because it was extremely difficult to gain access to HLA-B27$^+$ patients with *Chlamydia*-induced ReA, a disease becoming increasingly rare, or not unambiguously diagnosed (4), in Western countries. Attempts to stimulate peptide-specific, HLA-B27-restricted, CTL *in vitro* from a few individuals were unsuccessful. Due to the difficulties inherent to rising peptide-specific CTL *in vitro*, even from infected individuals, these studies must be performed with a sufficient number of patients, which was unfeasible, since they were not available. In the absence of formal confirmation with T cells, both the sequence homology and the predicted conformational features of DNAP(211-223) and B27(309-320) suggest a mechanism for increasing T-cell cross-reaction between endogenous chlamydial and self-derived HLA-B27 ligands through presentation of related peptides of distinct length and conformation, homologous to self-peptides with high flexibility in their bound state.

In conclusion, the high accuracy and sensitivity of current MS technologies brought about a major improvement in the detection of naturally processed HLA-B27 ligands from *C. trachomatis,* allowing us to detect 3 novel peptides from distinct proteins, including the second known HLA-B27-restricted epitope recognized by T cells from ReA patients. Both the homology of all the reported peptides with human sequences carrying the binding motif of HLA-B27 and the finding of a peptide from DNAP with significant sequence and conformational similarity to a human-derived HLA-B27 ligand, suggest that molecular mimicry between bacterial and self-derived HLA-B27 ligands may play a role in ReA. This mechanism could provide an autoimmune component that would exacerbate the pro-inflammatory role of HLA-B27, influencing disease severity and evolution toward chronicity.

Reference List

1. Brewerton, D. A., Hart, F. D., Nicholls, A., Caffrey, M., James, D. C., and Sturrock, R. D. (1973) *Lancet* **1,** 904-907

2. Schlosstein, L., Terasaki, P. I., Bluestone, R., and Pearson, C. M. (1973) *N.Engl.J.Med.* **288,** 704-706

3. Brewerton, D. A., Caffrey, M., Nicholls, A., Walters, D., Oates, J. K., and James, D. C. (1973) *Lancet* **2,** 996-998

4. Carter, J. D. and Hudson, A. P. (2010) *Curr.Opin.Rheumatol.* **22,** 424-430

5. Colmegna, I., Cuchacovich, R., and Espinoza, L. R. (2004) *Clin.Microbiol.Rev.* **17,** 348-369

6. Schiellerup, P., Krogfelt, K. A., and Locht, H. (2008) *J.Rheumatol.* **35,** 480-487

7. Kim, G. T. (2012) *Int.J.Rheum.Dis.* **15,** e113-e115

8. Beagley, K. W. and Timms, P. (2000) *J.Reprod.Immunol.* **48,** 47-68

9. Bachmaier, K. and Penninger, J. M. (2005) *Curr.Top.Microbiol.Immunol.* **296,** 153-163

10. Fan, T., Lu, H., Hu, H., Shi, L., McClarty, G. A., Nance, D. M., Greenberg, A. H., and Zhong, G. (1998) *J.Exp.Med.* **187,** 487-496

11. Belland, R. J., Scidmore, M. A., Crane, D. D., Hogan, D. M., Whitmire, W., McClarty, G., and Caldwell, H. D. (2001) *Proc.Natl.Acad.Sci.U.S.A* **98,** 13984-13989

12. Stenner-Liewen, F., Liewen, H., Zapata, J. M., Pawlowski, K., Godzik, A., and Reed, J. C. (2002) *J.Biol.Chem.* **277,** 9633-9636

13. Schwarzenbacher, R., Stenner-Liewen, F., Liewen, H., Robinson, H., Yuan, H., Bossy-Wetzel, E., Reed, J. C., and Liddington, R. C. (2004) *J.Biol.Chem.* **279,** 29320-29324

14. Jendro, M. C., Fingerle, F., Deutsch, T., Liese, A., Kohler, L., Kuipers, J. G., Raum, E., Martin, M., and Zeidler, H. (2004) *Med.Microbiol.Immunol.* **193,** 45-52

15. Fields, K. A. and Hackstadt, T. (2002) *Annu.Rev.Cell Dev.Biol.* **18,** 221-245

16. Beatty, W. L., Belanger, T. A., Desai, A. A., Morrison, R. P., and Byrne, G. I. (1994) *Infect.Immun.* **62,** 3705-3711

17. Beatty, W. L., Belanger, T. A., Desai, A. A., Morrison, R. P., and Byrne, G. I. (1994) *Ann.N.Y.Acad.Sci.* **730,** 304-306

18. Zhong, G., Fan, T., and Liu, L. (1999) *J.Exp.Med.* **189,** 1931-1938

19. Zhong, G., Liu, L., Fan, T., Fan, P., and Ji, H. (2000) *J.Exp.Med.* **191,** 1525-1534

20. Zhong, G., Fan, P., Ji, H., Dong, F., and Huang, Y. (2001) *J.Exp.Med.* **193,** 935-942

21. Zhong, G. (2011) *Front Microbiol.* **2,** 14

22. Loomis, W. P. and Starnbach, M. N. (2002) *Curr.Opin.Microbiol.* **5,** 87-91

23. Marcilla, M. and Lopez de Castro, J. A. (2008) *Tissue Antigens* **71,** 495-506

24. Benjamin, R. and Parham, P. (1990) *Immunol.Today* **11,** 137-142

25. Albert, L. J. and Inman, R. D. (1999) *N.Engl.J.Med.* **341,** 2068-2074

26. May, E., Dorris, M. L., Satumtira, N., Iqbal, I., Rehman, M. I., Lightfoot, E., and Taurog, J. D. (2003) *J.Immunol.* **170,** 1099-1105

27. Popov, I., Dela Cruz, C. S., Barber, B. H., Chiu, B., and Inman, R. D. (2001) *J.Immunol.* **167,** 3375-3382

28. Popov, I., Dela Cruz, C. S., Barber, B. H., Chiu, B., and Inman, R. D. (2002) *J.Immunol.* **169,** 4033-4038

29. Fourneau, J. M., Bach, J. M., Van Endert, P. M., and Bach, J. F. (2004) *Mol.Immunol.* **40,** 1095-1102

30. Bachmaier, K., Neu, N., de la Maza, L. M., Pal, S., Hessel, A., and Penninger, J. M. (1999) *Science* **283,** 1335-1339

31. Swanborg, R. H., Boros, D. L., Whittum-Hudson, J. A., and Hudson, A. P. (2006) *Expert.Rev.Mol.Med.* **8,** 1-23

32. Kuon, W., Holzhutter, H. G., Appel, H., Grolms, M., Kollnberger, S., Traeder, A., Henklein, P., Weiss, E., Thiel, A., Lauster, R., Bowness, P., Radbruch, A., Kloetzel, P. M., and Sieper, J. (2001) *J.Immunol.* **167,** 4738-4746

33. Appel, H., Kuon, W., Kuhne, M., Wu, P., Kuhlmann, S., Kollnberger, S., Thiel, A., Bowness, P., and Sieper, J. (2004) *Arthritis Res.Ther.* **6,** R521-R534

34. Wooldridge, L., Ekeruche-Makinde, J., van den Berg, H. A., Skowera, A., Miles, J. J., Tan, M. P., Dolton, G., Clement, M., Llewellyn-Lacey, S., Price, D. A., Peakman, M., and Sewell, A. K. (2012) *J.Biol.Chem.* **287,** 1168-1177

35. Karunakaran, K. P., Rey-Ladino, J., Stoynov, N., Berg, K., Shen, C., Jiang, X., Gabel, B. R., Yu, H., Foster, L. J., and Brunham, R. C. (2008) *J.Immunol.* **180,** 2459-2465

36. Yu, H., Jiang, X., Shen, C., Karunakaran, K. P., and Brunham, R. C. (2009) *J.Immunol.* **182,** 1602-1608

37. Ying, S., Fischer, S. F., Pettengill, M., Conte, D., Paschen, S. A., Ojcius, D. M., and Hacker, G. (2006) *Infect.Immun.* **74,** 6057-6066

38. Cragnolini, J. J. and Lopez de Castro, J. A. (2008) *Mol.Cell Proteomics* **7,** 170-180

39. Cragnolini, J. J., Garcia-Medel, N., and Lopez de Castro, J. A. (2009) *Mol.Cell Proteomics* **80,** 1850-1859

40. Calvo, V., Rojo, S., Lopez, D., Galocha, B., and Lopez de Castro, J. A. (1990) *J.Immunol.* **144,** 4038-4045

41. Long, E. O., Rosen-Bronson, S., Karp, D. R., Malnati, M., Sekaly, R. P., and Jaraquemada, D. (1991) *Hum.Immunol.* **31,** 229-235

42. Paradela, A., Garcia-Peydro, M., Vazquez, J., Rognan, D., and Lopez de Castro, J. A. (1998) *J.Immunol.* **161,** 5481-5490

43. Barnstable, C. J., Bodmer, W. F., Brown, G., Galfre, G., Milstein, C., Williams, A. F., and Ziegler, A. (1978) *Cell* **14,** 9-20

44. Paradela, A., Alvarez, I., Garcia-Peydro, M., Sesma, L., Ramos, M., Vazquez, J., and Lopez de Castro, J. A. (2000) *J.Immunol.* **164,** 329-337

45. Garcia-Medel, N., Sanz-Bravo, A., Barnea, E., Admon, A., and Lopez de Castro, J. A. (2012) *Mol.Cell Proteomics.* **11,** 1-15

46. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) *Electrophoresis* **20,** 3551-3567

47. Diez-Rivero, C. M., Lafuente, E. M., and Reche, P. A. (2010) *BMC.Bioinformatics.* **11,** 479

48. Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002) *Nucleic Acids Res.* **30,** 3059-3066

49. Hülsmeyer, M., Fiorillo, M. T., Bettosini, F., Sorrentino, R., Saenger, W., Ziegler, A., and Uchanska-Ziegler, B. (2004) *J.Exp.Med.* **199,** 271-281

50. Dolinsky, T. J., Czodrowski, P., Li, H., Nielsen, J. E., Jensen, J. H., Klebe, G., and Baker, N. A. (2007) *Nucleic Acids Res.* **35,** W522-W525

51. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006) *Proteins* **65,** 712-725

52. Case, D. A., Cheatham, T. E., III, Darden, T., Gohlke, H., Luo, R., Merz, K. M., Jr., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005) *J.Comput.Chem.* **26,** 1668-1688

53. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) *J.Chem.Phys.* **70,** 926-935

54. Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L., and Schulten, K. (2005) *J.Comput.Chem.* **26,** 1781-1802

55. Darden, T., York, D., and Pedersen, L. (1993) *J.Chem.Phys.* **98,** 10089-10092

56. Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001) *Proc.Natl.Acad.Sci.U.S.A* **98,** 10037-10041

57. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., and Baker, N. A. (2004) *Nucleic Acids Res.* **32,** W665-W667

58. Klett, J., Nuñez-Salgado, A., Dos Santos, H. G., Cortés-Cabrera, A., Perona, A., Gil-Redondo, R., Abia, D., Gago, F., and Morreale, A. (2012) *J.Chem.Theory Comput.* **8,** 3395-3408

59. R Development Core Team (2012) *R Foundation for Statistical Computing, Viena, Austria*

60. Dougan, D. A., Mogk, A., Zeth, K., Turgay, K., and Bukau, B. (2002) *FEBS Lett.* **529,** 6-10

61. Wang, F., Mei, Z., Qi, Y., Yan, C., Hu, Q., Wang, J., and Shi, Y. (2011) *Nature* **471,** 331-335

62. Ramos, M., Alvarez, I., Sesma, L., Logean, A., Rognan, D., and Lopez de Castro, J. A. (2002) *J.Biol.Chem.* **277,** 37573-37581

63. Lopez de Castro, J. A., Alvarez, I., Marcilla, M., Paradela, A., Ramos, M., Sesma, L., and Vazquez, M. (2004) *Tissue Antigens* **63,** 424-445

64. Ben Dror, L., Barnea, E., Beer, I., Mann, M., and Admon, A. (2010) *Arthritis Rheum.* **62,** 420-429

65. Starnbach, M. N., Loomis, W. P., Ovendale, P., Regan, D., Hess, B., Alderson, M. R., and Fling, S. P. (2003) *J.Immunol.* **171,** 4742-4749

66. Fields, K. A., Mead, D. J., Dooley, C. A., and Hackstadt, T. (2003) *Mol.Microbiol.* **48,** 671-683

67. Fields, K. A., Fischer, E. R., Mead, D. J., and Hackstadt, T. (2005) *J.Bacteriol.* **187,** 6466-6478

68. Chellas-Gery, B., Linton, C. N., and Fields, K. A. (2007) *Cell Microbiol.* **9,** 2417-2430

69. Amigorena, S. and Savina, A. (2010) *Curr.Opin.Immunol.* **22,** 109-117

70. Joffre, O. P., Segura, E., Savina, A., and Amigorena, S. (2012) *Nat.Rev.Immunol.* **12,** 557-569

71. Giles, D. K. and Wyrick, P. B. (2008) *Microbes.Infect.* **10,** 1494-1503

72. Purbhoo, M. A., Irvine, D. J., Huppa, J. B., and Davis, M. M. (2004) *Nat.Immunol.* **5,** 524-530

73. Belland, R. J., Zhong, G., Crane, D. D., Hogan, D., Sturdevant, D., Sharma, J., Beatty, W. L., and Caldwell, H. D. (2003) *Proc.Natl.Acad.Sci.U.S.A* **100,** 8478-8483

74. Saka, H. A., Thompson, J. W., Chen, Y. S., Kumar, Y., Dubois, L. G., Moseley, M. A., and Valdivia, R. H. (2011) *Mol.Microbiol.* **82,** 1185-1203

75. Belland, R. J., Nelson, D. E., Virok, D., Crane, D. D., Hogan, D., Sturdevant, D., Beatty, W. L., and Caldwell, H. D. (2003) *Proc.Natl.Acad.Sci.U.S.A* **100,** 15971-15976

76. Ruckert, C., Fiorillo, M. T., Loll, B., Moretti, R., Biesiadka, J., Saenger, W., Ziegler, A., Sorrentino, R., and Uchanska-Ziegler, B. (2005) *J.Biol.Chem.* **281,** 2306-2316

77. Nurzia, E., Panimolle, F., Cauli, A., Mathieu, A., Magnacca, A., Paladini, F., Sorrentino, R., and Fiorillo, M. T. (2010) *Clin.Immunol.* **135,** 476-482

Acknowledgements- We thank the staff of the Proteomics facilities at the Centro de Biología Molecular Severo Ochoa and Centro Nacional de Biotecnología, Madrid for help in MS.

FOOTNOTES

* This work was supported by grants SAF2008/00461, SAF2011/25681 (Plan Nacional de I+D+i), and RD08/0075 (RIER, ISCIII) to JALC, BSF #2009393 from the USA-Israel Binational Science Foundation to A.A, S2010-BMD-2457-BIPEDD2 (Comunidad Autónoma de Madrid) to A.M., who also acknowledges financial support from the AMAROUTO program (Fundación Severo Ochoa), and an institutional grant of the Fundación Ramón Areces to the CBMSO. C.A.N. is a fellow from the Ministry of Education of the Government of Chile. HGDS acknowledges financial support from grant BFU2011-24595 (Plan Nacional de I+D+i).

[1]The abbreviations used are: AS: ankylosing spondylitis; β2m: β2-microglobulin; CTL: cytolytic T lymphocyte; DNAP: DNA primase; EGFP: Enhanced GFP; C1R: Hmy2.C1R; HM: Homology modeling; MD: Molecular dynamics; MHC-I: MHC class I; PDB: Protein Data Bank; NQRA: (Na+)-translocating NADH-quinone reductase subunit A; PqqC: Pyrroloquinoline-quinone synthase-like protein; ReA: Reactive arthritis; Rep: Representative structure; RMSD: Root-mean square deviation; RMSF: Root-mean square fluctuation; RT: Retention time; SMIM: Selected multiple ion monitoring

FIGURE LEGENDS

FIGURE 1. Expression of ClpC fusion proteins in C1R-B*27:05 cells and search strategy for endogenous chlamydial peptides. *A,* schematic structure of ClpC and EGFP-ClpC fusion protein constructs. *B,* flow cytometry showing the EGFP-associated fluorescence of the indicated ClpC fusion protein transfectants. Untransfected C1R-B*27:05 cells (white) or cells transfected with EGFP alone (black) were included as controls. *C,* Western blot showing the stable expression of the indicated ClpC fusion proteins in the respective transfectant cells. The immunoblot was done on whole lysates with rabbit anti-GFP polyclonal antibody. *D,* experimental strategies for detecting chlamydial HLA-B27 ligands. The B*27:05-bound peptide pools from C1R-B*27:05 cells expressing or not the bacterial fusion protein were directly analyzed by LC-MS/MS using a LTQ-Orbitrap (1). Alternatively, a specific search was performed by determining the retention time (RT) of a target synthetic peptide (2) and analyzing the corresponding individual fractions, or a minipool of neighbor fractions around the RT of the synthetic peptide, from an HPLC-fractionated B27-bound peptide pool (3) and looking for the specific ion peaks at various charge states in a LTQ-Velos mass spectrometer (4). MS/MS spectra were submitted to automatic interpretation using the Mascot software (5). Each candidate sequence was revised manually and assisted by the MS-product tool (6). Final confirmation was done by comparing the MS/MS spectrum of the assigned peptide with that of the synthetic peptide (7).

FIGURE 2. Identification of the chlamydial B*27:05 ligand SRLDPVIGR from ClpC(1-512) transfectant cells. *A,* MS/MS spectra of the $[M+2H]^{2+}$ ion peaks at *m/z* 506.80 detected in the LTQ-Orbitrap from the unfractionated HLA-B27 peptidome (top), in the LTQ-Velos from Fr. 142 of the HPLC-fractionated HLA-B27 peptidome (middle) and the synthetic SRLDPVIGR peptide, corresponding to residues 203-211 of the ClpC protein (bottom). *B,* MS/MS spectrum of the $[M+3H]^{3+}$ ion peak at *m/z* 338.20 detected in a pool of HPLC fractions at the RT±3 min of the synthetic peptide, using an LTQ-Velos mass spectrometer (top), and of the synthetic peptide corresponding to residues 203-211 of the ClpC protein (bottom).

FIGURE 3. Identification of the chlamydial B*27:05 ligand mRDHTITLL from NQRA transfectant cells. MS/MS spectra of the $[M+2H]^{2+}$ ion peaks at *m/z* 558.33 detected in a LTQ-Velos mass spectrometer from a pool of fractions of the HPLC-fractionated B27 peptidome, corresponding to the RT±3 min of the synthetic peptide (top), and of the synthetic oxidized form of the sequence spanning residues 330-338 of the NQRA protein (bottom).

FIGURE 4. Identification of the chlamydial B*27:05 ligand RRFKEGGRGGKYI from DNAP transfectant cells. *A,* MS/MS spectra of the $[M+3H]^{3+}$ ion peaks at *m/z* 508.62 detected in the LTQ-Orbitrap from the unfractionated HLA-B27 peptidome (top), in a LTQ-Velos mass spectrometer from a pool of fractions of the HPLC-fractionated B27 peptidome, corresponding to the RT±3 min of the synthetic peptide middle), and of the synthetic peptide corresponding to residues 211-223 of the DNAP protein (bottom). *B,* MS/MS spectra of the $[M+2H]^{2+}$ ion peaks at *m/z* 762.43 detected in a pool of HPLC fractions at the RT ±3 min of the synthetic peptide, using an LTQ-Velos mass spectrometer (top), and of the synthetic peptide corresponding to residues 211-223 of the DNAP protein (bottom).

FIGURE 5. MD simulation of HLA-B*27:05 and β_2m and contribution of individual peptide residues to B*27:05 binding. *A,* Cα Root-Mean-Square deviation (RMSD, in Å) for each complex along the trajectories compared to their initial reference structures, HLA-B*27:05 heavy chain and β_2m are colored in blue and green, respectively. *B,* mass-weighted atomic positional fluctuations (RMSF, in Å) of the HLA-B27 heavy chain and β_2m for each HLA-B27/peptide complex, DNAP(211-221) (orange), DNAP(211-223) (brown), B27(309-320) (purple) and pVIPR-A (black). About 32% of the residues along the B*27:05 heavy chain, mainly in α3, showed RMSF values above 3.0Å. *C*, contribution of each single residue to the total free energy of binding of the corresponding peptide according to MM-ISMSA energy decomposition scheme: DNAP(211-221) (orange), DNAP(211-223) (red), B27(309-320) (blue) and pVIPR-A (black).

FIGURE 6. **Structural analysis of modeled HLA-B*27:05/peptide complexes.** *A,* Root-Mean-Square deviation (RMSD, in Å) corresponding to the peptidic Cα atoms along the MD trajectories, compared to their initial reference structures, for DNAP(211-221) (orange), DNAP(211-223) (brown), B27(309-320) (purple) and pVIPR-A (black). *B,* mass-weighted atomic positional fluctuations (RMSF, in Å) per residue for the four peptides (color code as in *A*). *C,* overlay of 100 structures sampled along the last 10 ns of the MD trajectories. The peptide, HLA-B*27:05 heavy chain (blue) and β2m (green) backbones are shown. *D,* molecular surface of representative peptide conformations (rep) for each of the main clusters obtained during the last 10 ns of MD simulation. O, N and other atoms are colored red, blue and white, respectively. *E,* Adaptive Poisson-Boltzmann solver (APBS) analysis for the most similar structures found during clustering. The distribution of electrostatic potentials on the peptide surfaces is shown. Negative and positive electrostatic potentials are colored red and blue, respectively (Range: ±5 Kcal).

Table I

Chlamydial HLA-B27 ligands processed in vivo from endogenous fusion proteins

Peptide	*Source Protein*	*Residues*	*Ref.*	*Predicted*[a]
KRALLEIVI	NQRA	86-94	(39)	No
MRDHTITLL	NQRA	330-338	This study	Yes
RRINREAERF	DNA Primase	112-121	(38)	N.A.
RRINREAERFF	DNA Primase	112-122	(38)	N.A.
RRFKEGGRGGK	DNA Primase	211-221	(38)	N.A.
RRFKEGGRGGKYI	DNA Primase	211-223	This study	N.A.
SRLDPVIGR	ClpC	203-211	This study	No
ARKLLLDNL	PqqC-like protein	70-78	(39)	Yes

[a]A combination of predictive binding and proteasome cleavage algorithms was used in a previous study to scan the proteome of *C. trachomatis* for potential HLA-B27-restricted nonameric epitopes, followed by antigen recognition assays *in vitro.* Only nonamers were searched (32). N.A.: not applicable.

Table II

Human sequences with high homology to chlamydial HLA-B27 ligands

Accession N.	Protein	Sequence[a]	Identity (%)	PCS[c]	ICS[c]
Sequences homologous to ClpC(203-211): SRLDPVIGR					
Q6V0I7	Protocadherin Fat 4	FRLDPVSGR	77	0.42	0.34
Q8TDY2	RB1-inducible coiled-coil protein 1	SRLDPRIIR	77	0.07	0.50
Q15493	Regucalcin	IRLDPVTGK	66	0.51	0.54
Q9UER7	Death domain-associated protein 6	SRLDEVISK	66	0.36	0.52
Q5VU43	Myomegalin	SRLEEVLGR	66	0.65	0.50
Q14406	Chorionic somatomammotropin hormone-like 1	SRLEPVRFL	55	0.77	0.68
Q8N3J3	Uncharacterized protein C17orf53	GRLRPVSSR	55	0.07	0.50
Q92935	Exostosin-like 1	LRLDPVLFK	55	0.17	0.52
Q16394	Exostosin-1	MRLDPVLFK	55	0.17	0.52
Q13753	Laminin subunit gamma-2	QRLDPVYFV	55	0.55	0.58
Q96BZ8	Leukocyte receptor cluster member 1	SRLDPLREM	55	0.55	0.60
Q86UR5	Regulating synaptic membrane exocytosis protein 1	SRLDPSSAV	55	0.53	0.58
Q6ZT12	E3 ubiquitin-protein ligase UBR3	SRLDPDYFI	55	0.61	0.55
Sequences homologous to NQRA(330-338): MRDHTITLL					
P48651	Phosphatidylserine synthase 1	YRPHTITLL[b]	77	0.67	0.64
P18510	Interleukin-1 receptor antagonist protein	LRSHLITLL	66	0.53	0.56
Q9Y2E8	Sodium/hydrogen exchanger 8	FRDHKITPK	55	0.23	0.53
Q2VPK5	Cytoplasmic tRNA 2-thiolation protein 2	MRDHTLKEV	55	0.55	0.58
Q6NSI8	Uncharacterized protein KIAA1841	VRDHMVTLR	55	0.50	0.51

[a]Identical residues to the bacterial sequences are underlined.
[b]This peptide contains Pro in P3, but it is shown here due to its homology with the bacterial ligand and high cleavage score.
[c]Constitutive proteasome (PCS) and immunoproteasome (ICS) cleavage scores (47). Values above 0.5 indicate high probability to generate the C-terminal end of the peptide.

Table III.
Clustering analysis for the indicated peptides.

	DNAP(211-221)		DNAP(211-223)		B27(309-320)		pVIPR-A	
Cluster #	**NS[a]**	**DC[b]**	**NS[a]**	**DC[b]**	**NS[a]**	**DC[b]**	**NS[a]**	**DC[b]**
1	734 (**14.7%**)	0.51	4987 (**99.7%**)	0.43	2473 (**49.5%**)	0.7	4984 (**99.7%**)	0.35
2	4193 (**83.9%**)	0.54	1 (0.0%)	0	559 (**11.2%**)	0.75	2 (0.0)	0.26
3	30 (0.6%)	0.43	1 (0.0%)	0	190 (3.8%)	0.67	3 (0.1%)	0.3
4	41 (0.8%)	0.4	3 (0.1%)	0.29	1777 (**35.5%**)	0.7	8 (0.2%)	0.3
5	2 (0.0%)	0.4	8 (0.2%)	0.31	1 (0.0%)	0	3 (0.1%)	0.19

[a] Number of structures. The percentages of the predominant clusters (in parentheses) are highlighted in bold
[b]Distance to centroid (Å).

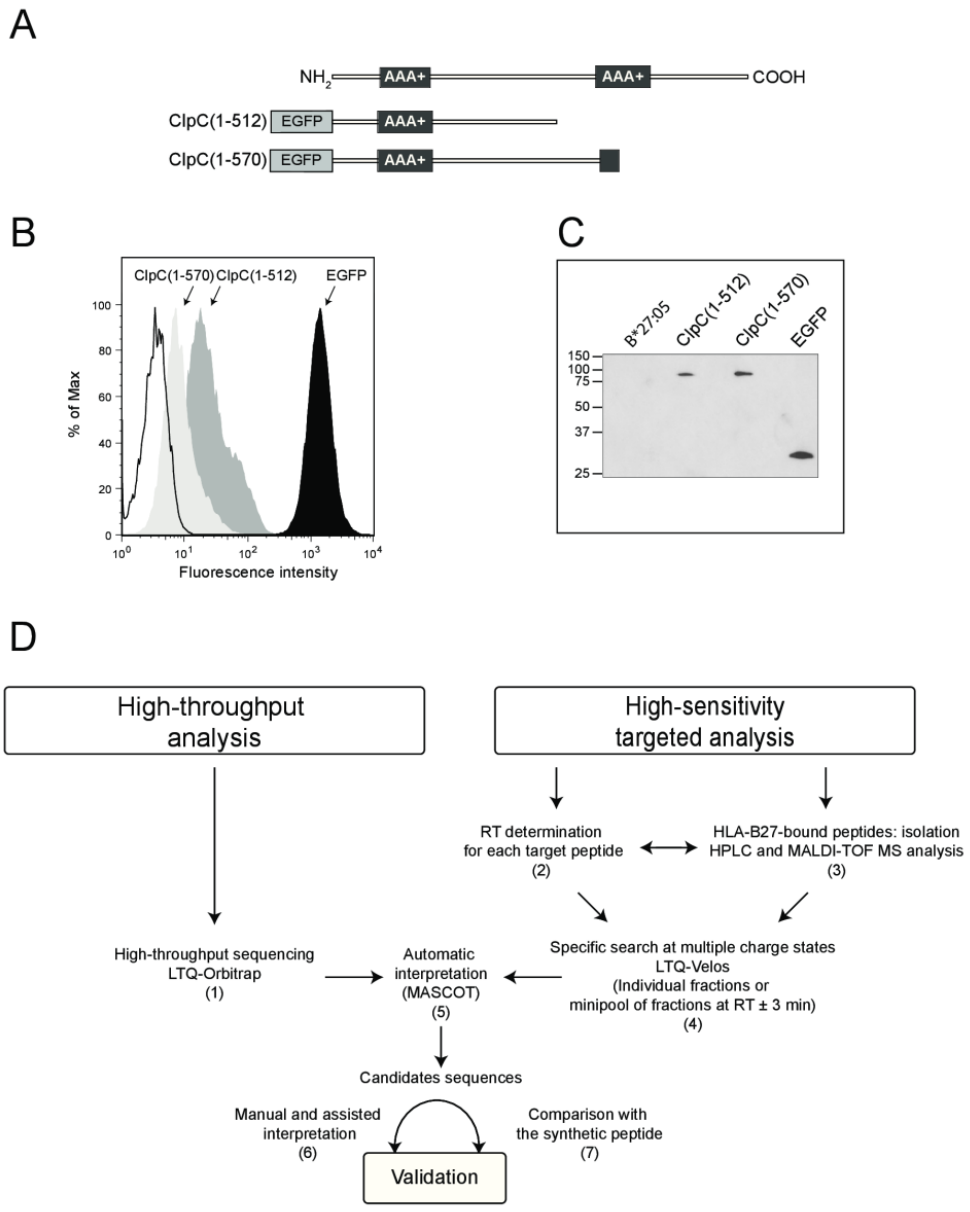


Figure 1

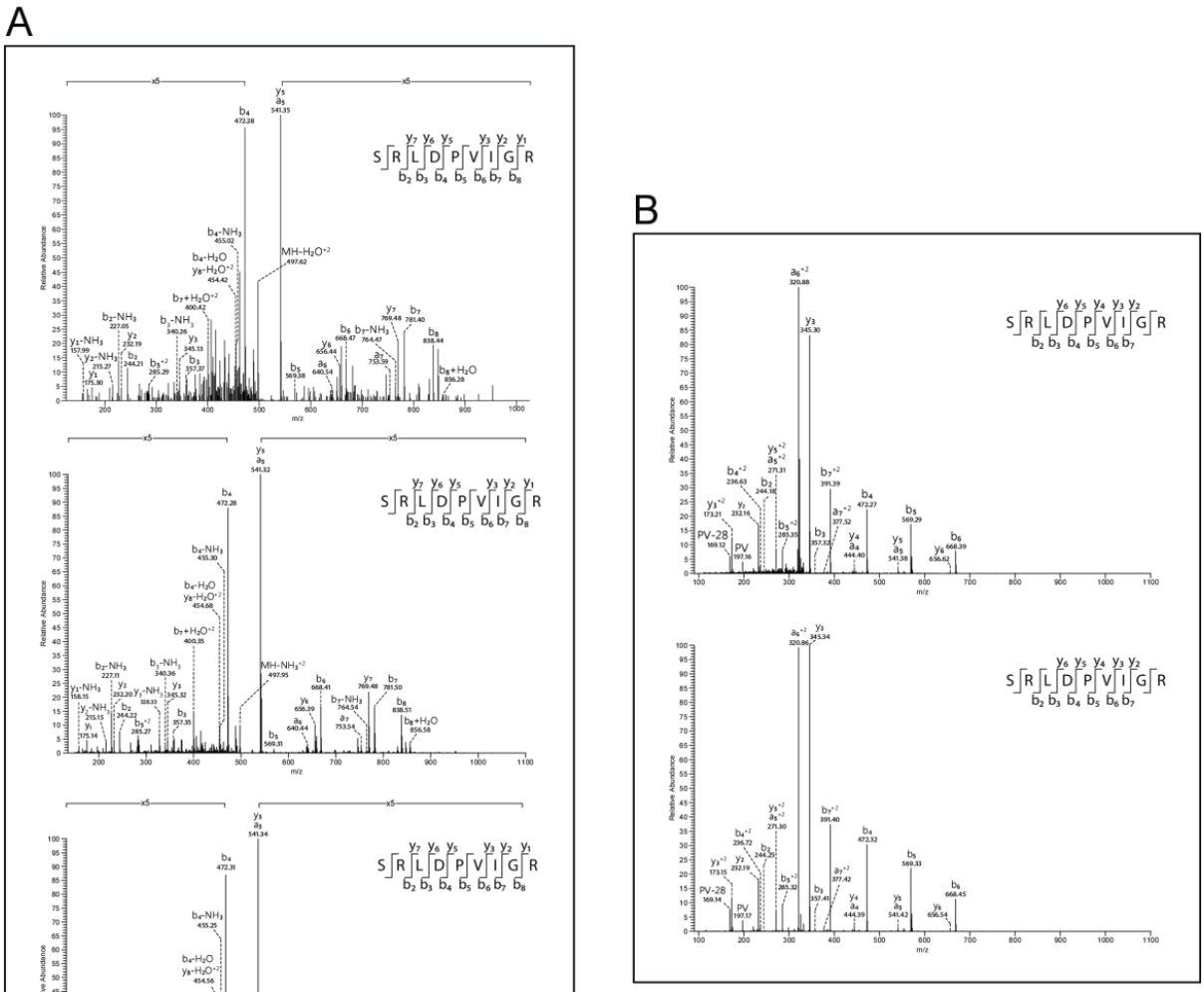


Figure 2

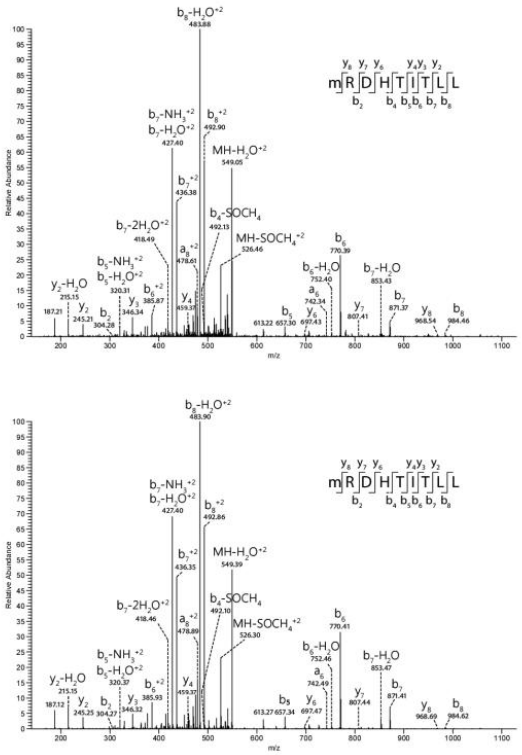
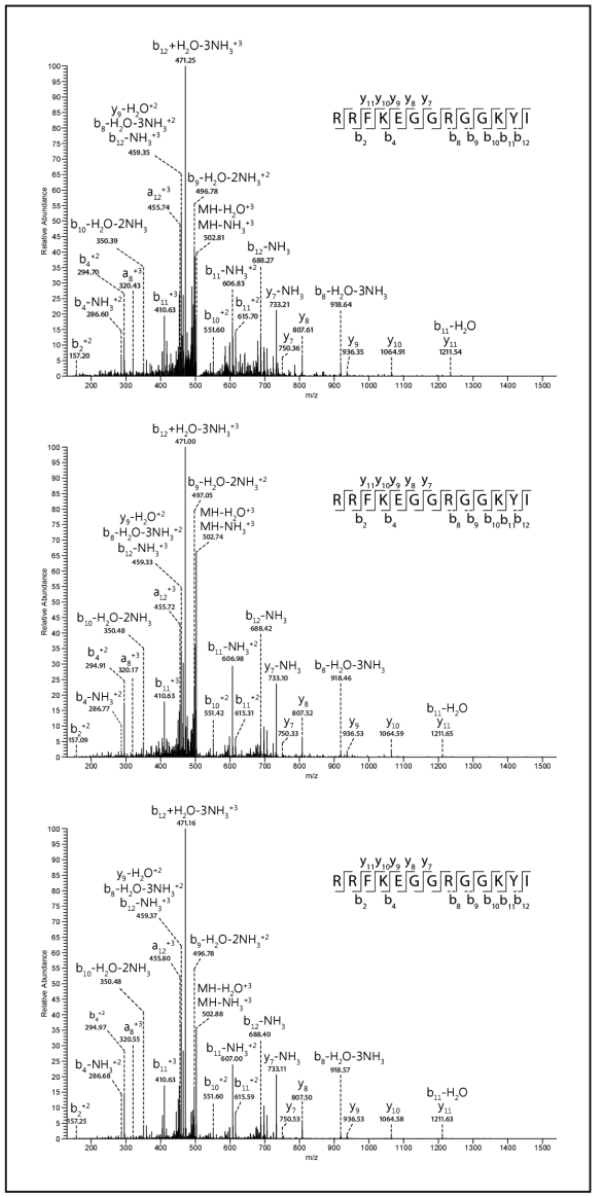


Figure 3

A



B

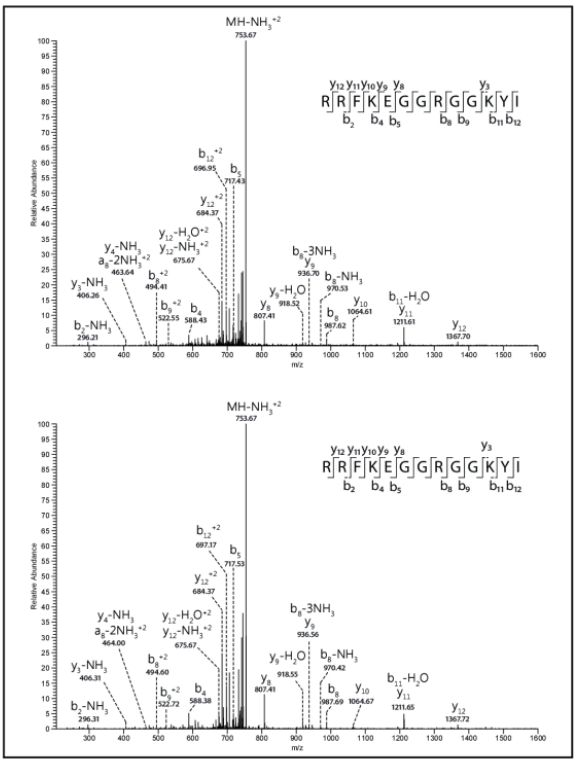


Figure 4

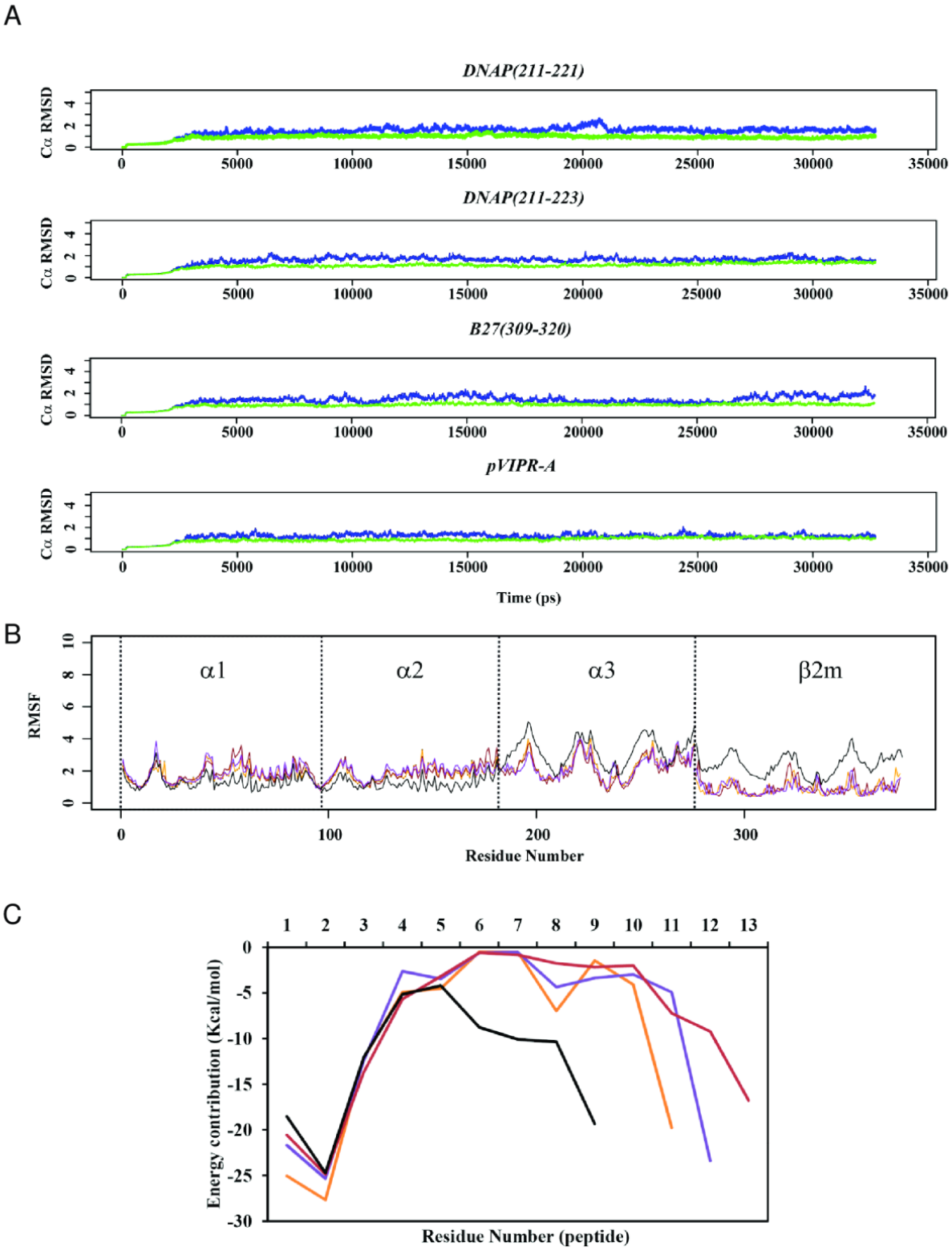


Figure 5

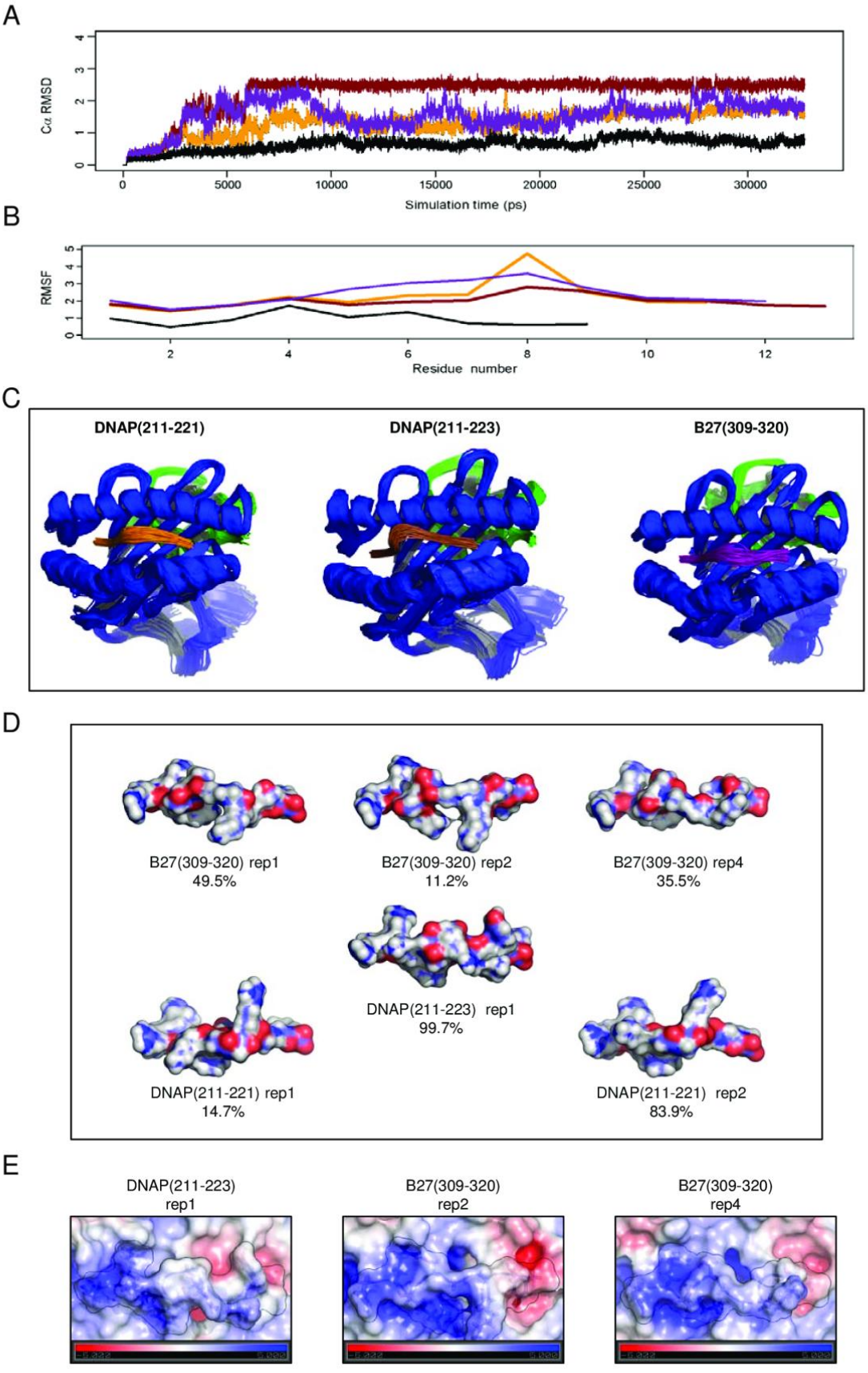


Figure 6

4.4.- Análisis masivo de la dinámica de los cambios conformacionales entre pares de estructuras cristalográficas de la base de datos PDB mediante modos normales torsionales

4.4.1.- Introducción y aportación del autor

La dinámica de las estructuras de las proteínas es esencial para su función, en particular durante la regulación alostérica y los cambios conformacionales, reacciones enzimáticas y el transporte de solutos a través de membranas biológicas. La escala temporal en la que ocurren los cambios conformacionales es muy amplia. Sin embargo, su dinámica térmica puede estudiarse analíticamente bajo la aproximación armónica usando la técnica de modos normales, que correlacionan con los cambios de conformación, el equilibrio dinámico así como con los cambios evolutivos. Los modos normales de red elástica (ENM), análogos a los modelos de GO, representan la mecánica estadística del proceso de plegamiento usando únicamente como datos de partida una estructura nativa conocida. Los modos de baja frecuencia describen las fluctuaciones colectivas en el estado nativo y poseen un solapamiento significativo con los cambios conformacionales experimentales. El éxito de los ENM indica que mucha de la información acerca de la flexibilidad de una proteína está contenida en su topología.

En una publicación reciente de nuestro laboratorio, Méndez y Bastolla (2011) presentaban un nuevo método para el cálculo de modos normales en el espacio torsional (TNM), un modelo de ENM que, considerando sólo los grados de libertad de los ángulos torsionales del *backbone* de la proteína, permite predecir sus movimientos naturales compatibles con su geometría covalente y topología. El TNM presenta 3 claras ventajas respecto a métodos anisotrópicos ANM: **(1)** promueven el desplazamiento preciso de más átomos sin coste computacional adicional, **(2)** la reducción de los grados de libertad de la proteína (de 3 por átomo a 2 por residuo si no consideramos cadenas laterales) por lo que podemos incluir las interacciones entre todos los átomos y **(3)** la conservación de la geometría de las estructuras generadas a nivel de distancias y ángulos de enlace y en menor medida a nivel de estructura secundaria, aplicando perturbaciones de pequeña amplitud.

En el presente trabajo (Dos Santos et al, 2013) nos hemos propuesto validar el método TNM a través del análisis masivo de cambios de conformación entre dos estructuras resueltas de la misma proteína. Hemos comparado estos cambios de conformación con un modelo nulo que hemos propuesto, basado en la teoría de la respuesta lineal, que predice que las proteínas son más deformables a lo largo de los modos de baja frecuencia, lo cual nos ha llevado a

predecir que, en un cambio conformacional al azar, las contribuciones de los modos normales a los cambios conformacionales (c_α^2) son proporcionales a sus fluctuaciones térmicas (ω^{-2}_α), o sea ($c_\alpha^2 \propto \omega^{-2}_\alpha$). Observamos además un exceso de correlación rho (ρ) con respecto al modelo nulo. Cuando $\rho > 0$, los modos normales de baja frecuencia contribuyen al cambio conformacional más de lo esperado por el modelo nulo, sólo pocos modos contribuyen, y las barreras de energía son pequeñas respecto a la amplitud del cambio estructural. Por lo tanto, interpretamos estas desviaciones del modelo nulo como un indicio de que la selección natural ha actuado sobre el movimiento natural de la proteína para reducir las barreras de energía que se oponen a sus movimientos funcionales.

La contribución de la autora de la tesis ha consistido en llevar a cabo un estudio masivo de todos los pares de estructuras monoméricas depositadas en el PDB con secuencia idéntica y con RMSD entre ellas > 1 Å. Se han distinguido diferentes tipos de pares: entre las formas apo (sin ligando unido) y holo (con ligando unido), entre pares que tienen los mismos ligandos, entre proteínas monoméricas y que son parte de complejos, entre la misma proteína fosforilada y sin fosforilar. En primer lugar, observamos que para todos los conjuntos el valor más frecuente del parámetro ρ es 0, lo que es de esperar en base al modelo nulo. Esto nos indica que los modos normales correlacionan con el cambio conformacional. Sin embargo no observamos esta correlación cuando el RMSD es muy pequeño (*i.e.* valores del orden del error experimental). Interpretamos estos resultados como una validación tanto del modelo nulo como del método TNM. Sin embargo, en algunos conjuntos observamos un gran número de pares con valor significativo de $\rho > 0$, esperables para movimientos funcionales tales como reacciones enzimáticas y de transporte. Estos pares son particularmente frecuentes en el conjunto de complejos con múltiples cadenas y en el conjunto de proteínas fosforiladas. En este conjunto son también frecuentes casos con $\rho < 0$ con barreras de energía grandes respecto a la amplitud del cambio, como esperábamos encontrar en cambios de conformación involucrados en procesos de regulación dado que éstos no deben que ocurrir de forma espontánea.

En conclusión, los resultados obtenidos no sólo han permitido la validación del método TNM y nuestro modelo nulo de cambios de conformación sino también proponen un método para identificar, a través del valor de ρ, los cambios de conformación con posible valor funcional. Finalmente, este marco unifica dos de los escenarios propuestos para la unión molecular (ver sección 1.2.1. de modelos de unión): la selección conformacional entre los estados pre-existentes, que en este marco interpretamos como la existencia de una

correlación entre la dinámica intrínseca de la proteína y la deformación producida por la unión del ligando ($\rho > 0$) y el ajuste inducido, que en este marco interpretamos como lo que se espera bajo respuesta lineal, es decir el modelo nulo ($\rho \approx 0$).

Artículo 4

Biochimica et Biophysica Acta 1834 (2013) 836–846

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbapap

Characterizing conformation changes in proteins through the torsional elastic response ☆

Helena G. Dos Santos [1], Javier Klett [1], Raúl Méndez, Ugo Bastolla *

Centro de Biologia Molecular Severo Ochoa, CSIC-UAM, 28049 Madrid, Spain

ARTICLE INFO

Article history:
Received 29 November 2012
Received in revised form 22 January 2013
Accepted 6 February 2013
Available online 19 February 2013

Keywords:
Normal mode analysis
Elastic network model
Torsion angle
Conformation change
Allostery
Energy barrier

ABSTRACT

The relationship between functional conformation changes and thermal dynamics of proteins is investigated with the help of the torsional network model (TNM), an elastic network model in torsion angle space that we recently introduced. We propose and test a null-model of "random" conformation changes that assumes that the contributions of normal modes to conformation changes are proportional to their contributions to thermal fluctuations. Deviations from this null model are generally small. When they are large and significant, they consist in conformation changes that are represented by very few low frequency normal modes and overcome small energy barriers. We interpret these features as the result of natural selection favoring the intrinsic protein dynamics consistent with functional conformation changes. These "selected" conformation changes are more frequently associated to ligand binding, and in particular phosphorylation, than to pairs of conformations with the same ligands. This deep relationship between the thermal dynamics of a protein, represented by its normal modes, and its functional dynamics can reconcile in a unique framework the two models of conformation changes, conformational selection and induced fit. The program TNM that computes torsional normal modes and analyzes conformation changes is available upon request. This article is part of a Special Issue entitled: The emerging dynamic view of proteins: Protein plasticity in allostery, evolution and self-assembly.

1. Introduction

Proteins are molecular machines that perform their biological function dynamically [1,2]. Stability, i.e., the existence of a well-defined average three-dimensional structure, and at the same time flexibility, i.e., the existence of intrinsic collective movements of large amplitude, characterize the native state of ordered proteins and are key for their catalytic activity [3], ligand binding ability [4], and allosteric regulation [5].

For ordered proteins, the topology of the native state determines to a large extent both the stability of the protein and its intrinsic collective dynamics, which can be predicted by elastic network models (ENM) [6–8,12]. They are Go-like models [9,10] that represent the energetics of the native state based only on its topology. Go models and ENMs fulfill the principle of minimal frustration [11], which assumes that all native interactions are at their energy minimum. Several flavors of ENM have been described in the literature [13]. We have recently introduced the torsional network model (TNM) [14], which adopts the torsion angles of the protein backbone as degrees of freedom, similar to other methods of normal mode analysis and protein dynamics in torsion angle space [15–19]. The TNM has the advantage that it represents all protein atoms with a computationally affordable cost, concentrating on physically allowed motions that do not modify bond lengths and bond angles.

The intrinsic dynamics of the native state of a protein modeled through the ENM can be analytically studied using normal mode analysis (NMA) [20]. NMA approximates the native energy landscape as the harmonic well in the neighborhood of the equilibrium position, and decomposes the native ensemble into a set of independent motions, the normal modes. Low frequency normal modes tend to represent collective motions that produce the largest displacements from the average position. It has been observed that the low frequency normal modes of the ENMs correlate with the intrinsic motions of the protein measured by crystallographic B-factors [15,21,22], despite the fact that B-factors are largely influenced by rigid body motions not represented by NMA [23], they correlate with the essential motions produced by very long molecular dynamics simulations [24,25], and with functional conformation changes such as those upon binding of a physiological ligand, in the sense that often a few low frequency normal modes almost perfectly reproduce the functional motion [26–29].

In this work we investigate the relationship between intrinsic protein motions predicted through ENMs and protein motions observed

☆ This article is part of a Special Issue entitled: The emerging dynamic view of proteins: Protein plasticity in allostery, evolution and self-assembly.

* Corresponding author. Tel.: +34 911964633.
E-mail address: ubastolla@cbm.uam.es (U. Bastolla).

[1] These authors contributed equally to this work.

1570-9639/$ – see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.bbapap.2013.02.010

as two different conformations of the same protein determined in X-ray crystallography experiments or NMR spectroscopy. Linear response theory predicts that the response of the protein to a generic perturbation, for instance ligand binding, is mostly influenced by low frequency normal modes [30,31]. Based on linear response theory, we recently proposed a null model of the response to a "random" perturbation that is independent of the intrinsic dynamics of the proteins, and introduced the parameter ρ that quantifies significant deviations from this null model [14]. When this parameter is large, low frequency normal modes contribute to the conformation change significantly more than expected based on the null model, and the energy barrier opposing to the conformation change is reduced. We observe that in this case only a small number of low frequency normal modes are sufficient to reproduce the conformation change. We interpret these observations as hints of a co-evolution between the functional motion and the intrinsic dynamics of the protein. Here we perform a large scale study of the relationship between conformation changes and intrinsic protein dynamics, analyzing all pairs of crystallized protein structures having the same amino acid sequence and at least 1 Å of root mean square deviation (RMSD).

The paper is organized as follows. In the first section we review the torsional network model used in the computations, reporting computational details omitted in the original publication. In the second section we present our null model of conformation change and the parameter ρ that measures deviations from the null model. In the third section we present the results of a massive analysis of the protein data bank.

2. Torsional network model (TNM)

2.1. Degrees of freedom and kinematics

A molecule composed of n atoms with masses m_i can be represented either through their Cartesian coordinates $\left\{\vec{r}_i\right\}$ or, equivalently, through a set of $n-1$ bonds connecting pairs of atoms, each characterized by the bond length l_a, the bond angle θ_a that it forms with the previous bond and the torsion angle φ_a with respect to the plane of the two previous bonds. In the TNM, only the backbone torsion angles phi (rotation around to the N–C_α bond) and psi (rotation around the C_α–C bond) are allowed to vary, and the other degrees of freedom are kept fixed. Our computer program allows us to additionally select the backbone angle omega and side-chain torsion angles as degrees of freedom, but usually this introduces noise and considerably increases the computation time. We then have to select reference atoms for computing kinetic energy. These can be only α carbons, only β carbons, all backbone atoms, all backbone atoms plus β carbons, or all heavy atoms, which are the best choice, on which the results presented in this paper are based. Bond lengths and bond angles are treated as degrees of freedom if the bond is not a covalent bond, such as the virtual bond connecting two residues separated by a disordered loop whose coordinates cannot be determined in the X-ray experiment. When we analyze a conformation change, we consider for the computation of kinetic energy only residues that are aligned in the two structures and treat gaps in the alignment as disordered loops, nevertheless all atoms are used for computing native interactions (see below). We denote with d the number of degrees of freedom in the torsional space.

The Jacobian matrix that relates infinitesimal torsional and Cartesian displacements is

$$\vec{J}_{ia} \equiv \frac{\partial \vec{r}_i}{\partial \varphi_a} = \chi_{ia}\left(\vec{\tau}_a + \vec{v}_a \times \vec{r}_i\right) \tag{1}$$

where $\chi_{ia} \in \{0,1\}$ is one if atom i is upstream of axes a and zero otherwise (by convention, torsional perturbations are propagated from the N-terminus to the C-terminus of the protein), $\times$ denotes the vector product and $\vec{\tau}_a$ and $\vec{v}_a$ are the translation and the rotation associated with the degree of freedom a, respectively. If the degree of freedom represents the torsion around the axis with unit vector $\vec{e}_a$ and origin $\vec{s}_a$, it is easy to see that $\vec{v}_a = \vec{e}_a$ and $\vec{\tau}_a = -\vec{e}_a \times \vec{s}_a$. If the degree of freedom represents the bond length, it holds $\vec{v}_a = 0$ and $\vec{\tau}_a = \vec{e}_a$ and if it represents the bond angle, it holds $\vec{v}_a = \left(\vec{e}_{a-1} \times \vec{e}_a\right) / \left|\vec{e}_{a-1} \times \vec{e}_a\right|$ and $\vec{\tau}_a = -\vec{v}_a \times \vec{s}_a$.

The degree of freedom a modifies both the internal degrees of freedom and the rigid body degrees of freedom. To get rid of the latter, we have to impose the Eckart conditions [32]

$$\sum_i m_i \vec{J}'_{ia} = 0, \sum_i m_i \vec{r}_i \times \vec{J}'_{ia} = 0 \tag{2}$$

resulting in $\vec{J}'_{ia} = \chi_{ia}\left(\vec{\tau}_a + \vec{v}_a \times \vec{r}_i\right) + \left(\vec{\tau}'_a + \vec{v}'_a \times \vec{r}_i\right)$, i.e., we have to apply the rigid body transformations $\vec{\tau}'_a$ and $\vec{v}'_a$ that compensate the rigid body motion of the molecule. In the reference frame in which the center of mass is at the origin, it holds $\vec{\tau}'_a = -\frac{M_a}{M}\left(\vec{\tau}_a + \vec{v}_a \times \vec{R}_a\right), I\vec{v}'_a = -I_a\vec{v}'_a - M_a\vec{R}_a \times \vec{\tau}_a$, with $M = \sum_i mi$, $M_a = \sum_i \chi_{ia} m_i$, $M_a\vec{R}_a = \sum_i X_{ia} m_i \vec{r}_i$, I is the inertia tensor and I_α is its restriction to the set having $\chi_{ia} = 1$ [15].

The kinetic energy matrix T in torsion angle space is

$$T_{ab} = \sum_i m_i \frac{\partial \vec{r}_i}{\partial \varphi_a}\frac{\partial \vec{r}_i}{\partial \varphi_b} = \sum_i m_i \vec{J}'_{ia}\vec{J}'_{ib} \times \tag{3}$$

In matrix notation, $T = J'^t M J' = K^t K$, where J' is represented as a $3n \times d$ matrix, the superscript t indicates matrix transposition, M is the diagonal mass matrix (not to be confused with the total mass $M = \sum_i m_i$), and we introduce the notation $K_{ia} = \sqrt{m_i}\vec{J}'_{ia}$. Taking advantage of the Eckart conditions, we can simplify the formula as

$$\begin{aligned} T_{ab} = {} & M_{ab}\vec{\tau}_a \cdot \vec{\tau}_b + \vec{v}_a \cdot I^{ab}\vec{v}_b \\ & + M_{ab}\vec{R}_{ab} \cdot \left(\vec{\tau}_a \times \vec{v}_b + \vec{\tau}_b \times \vec{v}_a\right) - M\vec{\tau}'_a \cdot \vec{\tau}'_b - \vec{v}'_a \cdot I\vec{v}'_b. \end{aligned} \tag{4}$$

Here M_{ab}, $\vec{R}_{ab}$ and I_{ab} are the mass, center of mass and inertia tensor of the set of atoms that are moved by both degrees of freedom a and b, i.e., $\chi_{ia} = \chi_{ib} = 1$. We now exploit the fact that the degrees of freedom are nested, i.e., if axis b is downstream of axis a (which we denote as $b > a$), then $\chi_{ib} = 1$ implies $\chi_{ia} = 1$, so that $M_{ab} = M_b$, unless a and b represent degrees of freedom of different side-chains, in which case $M_{ab} = 0$.

2.2. Potential energy

In ENMs, the effective potential energy of the protein is modeled as a sum of pairwise terms that only runs over native interactions, $V = \sum_{ij} C_{ij} v(r_{ij})$. $r_{ij} = \left|\vec{r}_i - \vec{r}_j\right|$ is the distance between interacting atoms. $C_{ij} = 1$ if the atom i and j are in contact in the native state, 0 otherwise. We use a definition of contacts in which for each pair of residues the two heavy atoms at shorter distance interact provided that their distance is smaller than 4.5 Å. The Go model (or, equivalently, the principle of minimum frustration) requires that each interaction term has a minimum corresponding to the native interaction distance r^0_{ij}. For small displacements from the equilibrium position $\left\{\vec{r}^{\,0}_i\right\}$, the potential energy can be expanded in Taylor series up to second order. Since the constant term can be ignored, and the force

at equilibrium vanishes, we only have to consider the quadratic form of the Hessian matrix,

$$V_0 \approx \frac{1}{2}\sum_{ij} C_{ij} f\left(r_{ij}^0\right)\left(r_{ij}-r_{ij}^0\right)^2. \quad (5)$$

We adopt a force constant that decays with distance as $f(r)=r^{-2}$, which has been proposed to improve the performances of the ENM [22]. The Hessian matrix in torsion angle space is

$$H_{ab}^{(\varphi)} \equiv \frac{\partial^2 V_0}{\partial\varphi_a \partial\varphi_b} = \sum_{ij} \frac{\partial^2 V_0}{\partial \vec{r}_i \partial \vec{r}_j} \frac{\partial \vec{r}_i}{\partial\varphi_a} \frac{\partial \vec{r}_j}{\partial\varphi_b}. \quad (6)$$

In matrix notation, $H^{(\varphi)}=J^t H^{(r)} J$, where $H^{(r)}$ is the Hessian in Cartesian coordinates [8]. We use the Jacobian matrix $J_{ia}=\chi_{ia}\left(\vec{v}_a\times\vec{r}_i+\vec{\tau}_a\right)$ instead of J' that subtracts rigid body displacements, since they do not modify the interatomic distances. Note that the terms in the sum are zero if both atoms i and j are modified by the torsion angle a (or b), or if none of them is modified by a (or b). The only non-zero terms are those in which one interacting atom j is upstream to both axes a and b and the other atom $i>j$ is downstream to both axes. Taking this into account, we can simplify the above computation as

$$H_{ab}^{(\varphi)} = \sum_{i>a\geq b\geq j} f\left(r_{ij}^0\right) C_{ij} \frac{\vec{r}_{ij}}{\left|\vec{r}_{ij}\right|}\left(\vec{v}_a\times\vec{r}_i+\vec{\tau}_a\right)\frac{\vec{r}_{ij}}{\left|\vec{r}_{ij}\right|}\left(\vec{v}_b\times\vec{r}_i+\vec{\tau}_b\right) \quad (7)$$

where $i>a$ means that atom i is downstream of axis a.

2.3. Normal mode analysis

For small perturbations within the harmonic approximation, the Lagrangian of the molecule $\mathscr{L}=T-V$ is a quadratic function and the solution of the dynamical equations can be represented as the sum of normal modes, which undergo periodic fluctuations of frequency ω_α, $\Phi_a(t)=A_\alpha u_a^\alpha e^{i\omega_\alpha t}$. The normal mode coordinates u_a^α satisfy the Lagrangian matrix equation

$$\left(J^t H^{(r)} J\right)u^\alpha = \omega_\alpha^2 T u^\alpha = \omega_\alpha^2 L^t L u^\alpha, \quad (8)$$

where we have written the symmetric and positive matrix T as $T=L^tL$. The matrix L can be easily obtained for instance through Cholesky decomposition, which has the computational advantage to return a lower tridiagonal matrix titL. The standard way to solve the above equation consists in introducing the mass-weighted normal mode coordinates $v^\alpha=Lu^\alpha$ that satisfy the eigenvalue equation

$$\left(JL^{-1}\right)H^{(r)}\left(JL^{-1}\right)v^\alpha \equiv \tilde{H}^{(\phi)} = \omega_\alpha^2 v^\alpha, \quad (9)$$

which can be solved by matrix diagonalization.

From the computational point of view, we compute the matrix $\tilde{\vec{J}}_{ib}=\sum_{a>i}\vec{J}_{ib}L_{ab}^{-1}=\sum_{a>i}\vec{v}_a L_{ab}^{-1}\times\vec{r}_i+\sum_{a>i}\vec{\tau}_a L_{ab}^{-1}=\tilde{\vec{v}}_{a(i)b}\times\vec{r}+\tilde{\vec{\tau}}_{a(i)b}$, with $\tilde{\vec{v}}_{a(i)b}=\sum_{a>a(i)}\vec{v}_a L_{ab}^{-1}$ and $\tilde{\vec{\tau}}_{a(i)b}=\sum_{a>a(i)}\vec{\tau}L_{ab}^{-1}$, where $a(i)$ is the most upstream degree of freedom such that $\chi_i=1$. Note that the dimension of the matrices $\tilde{\vec{\tau}}_{ib}$ and $\tilde{\vec{v}}_{ib}$ is only $3\times d^2$ and not $3\times d\times 3n$.

From the mass-weighted normal modes v^α, we obtain the torsional normal modes as $u^\alpha=L^{-1}v^\alpha$ and the Cartesian normal modes as $x^\alpha=J'L^{-1}v^\alpha$. Note that the mass-weighted normal modes are orthonormal, since they are the eigenvectors of a symmetric matrix: $\langle v^\alpha,v^\beta\rangle=\delta_{\alpha\beta}$, which implies that the Cartesian and torsional normal modes are orthonormal with respect to the scalar product weighted by the kinetic energy tensor, a general property of normal mode analysis: $(x^\alpha,Mx^\beta)=(J'u^\alpha,MJ'u^\beta)=\langle u^\alpha,(J')^tMJ'u^\beta\rangle=\langle u^\alpha,Tu^\beta\rangle=\langle Lu^\alpha,Lu^\beta\rangle=\langle v^\alpha,v^\beta\rangle=\delta_{\alpha\beta}$. Here and in the following, we denote with (X, Y) the scalar product of the $3n$-dimensional vectors X and Y in Cartesian space, with $\langle X,Y\rangle$ the scalar product of the d-dimensional vectors X and Y in torsional space, and with $\langle X\rangle$ the thermal average of X.

According to the equipartition theorem, each normal mode contributes equally to the average kinetic energy of the ensemble, which implies that low frequency normal modes determine large root mean square displacements from the equilibrium position. The contribution of mode α to thermal fluctuations is thus $\sum_i mi\langle\delta\vec{r}_i^2\rangle_\alpha=(k_BT/\omega_\alpha^2)\sum_i m_i\left|\vec{x}_i^\alpha\right|^2\propto\omega_\alpha^{-2}$.

2.4. Building deformed structures with the build-up algorithm

The Cartesian torsional modes preserve bond lengths and bond angles only up to first order in the amplitudes A_α. For larger deformations, in order to preserve bond angles and bond lengths, we construct the deformed structure one atom after the other using as coordinates bond angles, bond lengths and torsion angles [33]. The degrees of freedom modeled in the TNM are incremented by the normal coordinates with respect to their values in the unperturbed structure φ_a^0:

$$\varphi_a = \varphi_a^0 + A_\alpha u_a^\alpha. \quad (10)$$

Since finite rotations do not commute, the final structure will be different if we construct the molecule from the N-terminus to the C-terminus or vice versa. We choose the convention to build the molecule from the N-terminus to the C-terminus.

3. Analysis of conformation changes

3.1. Data sets

We selected all pairs of structures in the PDB with identical sequences, structures determined by X-ray crystallography, and RMSD greater than 1 Å in order to eliminate pairs for which the conformation change is of a similar order as the experimental error for structure determination. For each pair of proteins A and B, we examined the conformation changes from A to B and from B to A. We separately examined the following data sets:

1. Monomeric structures with the same ligands in the two conformations, crystallized in different crystals (different PDB codes): 6230 pairs.
2. Monomeric structures with different ligands in the two conformations such as apo and holo conformations of the same enzyme: 87,230 pairs.
3. Homodimeric structures with the same ligands in the two chains, crystallized in the same crystal (same PDB codes): 17,622 pairs.
4. Pairs of unphosphorylated–phosphorylated proteins, in which the "ligand" is a covalently bound phosphate group: 12,100 pairs.

We also considered the NMR ensemble of the protein ubiquitin (116 structures, PDB code 2k39) and compared it with the normal modes of the crystal structure of unbound ubiquitin (PDB code 1ubi).

3.2. Torsional displacements and torsional fraction

We compare two conformations A and B of the same protein, for instance the open and closed forms of an enzyme, or the phosphorylated and unphosphorylated forms of a protein, but also non functional conformation changes, such as the same protein crystallized in different conditions, or two chains of a homo-dimer. First, we align protein sequences to take care of possible differences in the crystallized constructs or in disordered regions without electronic density. Only aligned atoms that are present in both structures are used as reference atoms for computing the kinetic energy, but all atoms of the unperturbed structure A are used to compute the contact matrix and the Hessian matrix. Aligned atoms are superimposed minimizing the

mass-weighted deviation $\sum_i m_i \left|\vec{r}_i^{\,A} - \vec{r}_i\, B\right|^2$. This imposes the Eckart conditions $\sum_i m_i \left(\vec{r}_i^{\,A} - \vec{r}_i\, B\right) = 0$, $\sum_i m_i\, \vec{r}_i\, A \times \left(\vec{r}_i^{\,A} - \vec{r}_i\, B\right) = 0$, so that rigid body displacements do not contribute to the conformation change.

The torsional deviations are estimated as the angles whose associated Cartesian displacements $J\Delta\varphi^{AB}$ best fit the observed $\Delta\vec{r}_i^{\,AB} \equiv \vec{r}_i^{\,A} - \vec{r}_i^{\,B}$, i.e., we minimize $\sum_i m_i \left(\Delta\vec{r}_i - \sum_a \vec{J}_{ia}\Delta\varphi_a\right)$ by linear least squares, obtaining

$$\Delta\varphi = \left(J^t M J\right)^{-1} J^t M \Delta r \tag{11}$$

The matrix $J^tMJ = T$ is the kinetic energy, Eq. (3). We define the torsional fraction of a Cartesian conformation change as $(J\Delta\varphi, MJ\Delta\varphi)/(\Delta r, M\Delta r)$. A torsional fraction smaller than one means that some motions not considered in the TNM, such as changes in bond angles, bond lengths, omega torsion angles or side-chain torsion angles, contribute significantly to the conformation change.

The projections of the conformation change onto the normal modes of the unbound structure are computed as:

$$c_\alpha = \left(\Delta r^{AB}, Mx^\alpha\right) = \left\langle \Delta\varphi^{AB}, Tu^\alpha \right\rangle = \left\langle L\Delta\varphi^{AB}, v^\alpha \right\rangle. \tag{12}$$

The projection on the normal modes is the same using torsional or Cartesian displacements, since $\langle \Delta\varphi, Tu^\alpha\rangle = \langle T\Delta\varphi, u^\alpha\rangle = \langle J^tM\Delta r, u^\alpha\rangle = (\Delta r, Mx^\alpha)$. The sum of the squared projections is proportional to the torsional fraction, since the degrees of freedom that are not represented in the TNM have zero projection on TNM normal modes.

3.3. Collectivity of the conformation change

The collectivity of the conformation change measures the number of degrees of freedom that significantly contribute to it. Following Ref. [34], we measure the fractional contribution of each Cartesian degree of freedom, $p_i = m_i\Delta r_i^2 / \sum_j m_j \Delta r_j^2$ and measure the Cartesian collectivity as the exponential of the Shannon entropy of p_i, $\kappa[\{p_i\}] = \exp(-\sum_i p_i \log p_i)$, which can be roughly interpreted as the number of degrees of freedom with significant weight p_i. Similarly, the mass-weighted torsional collectivity is obtained from $p_a = (L\Delta\varphi)_a^2 / \sum_b (L\Delta\varphi)_b^2$, and the torsional collectivity without weights is $p_a = \Delta\varphi_a^2 / \sum_b \Delta\varphi_b^2$. A conformation change, or a normal mode, with large Cartesian collectivity or mass-weighted torsional collectivity is collective in the sense that it moves many atoms. If a conformation change or a normal mode has small unweighted torsional collectivity and large Cartesian collectivity, it means that it moves few torsion angles, as one would expect for a hinge motion [35].

The reciprocal collectivity of the conformation change is the collectivity of $p_\alpha = c_\alpha^2 / \sum_\beta c_\beta^2$. If the reciprocal collectivity of the conformation change is small, only a small number of normal modes contribute significantly to it. We normalize the reciprocal collectivity by the number of normal modes, so that it represents the fraction of normal modes that contribute to the conformation change.

3.4. Harmonic energy barriers

Within the harmonic approximation, the potential energy corresponding to a deformation of ΔE^{AB} in the direction of mode α is

$$\Delta E^{AB} = \frac{1}{2}\sum_\alpha (c_\alpha \omega_\alpha)^2 \Theta\left(|c_\alpha|/\sqrt{M} - \epsilon\right). \tag{13}$$

The theta function filters out high frequency modes that produce displacements that are smaller than the experimental resolution, in particular we require that the displacement produced by each mode α is larger than $=0.1$ Å.

3.5. Torsional linear response

If the harmonic potential V_0 of the ENM is perturbed by adding a potential V, for instance due to ligand binding, linear response theory in Cartesian space predicts that the equilibrium positions are displaced by $\Delta r = -\left(H^{(r)}\right)^{-1} \frac{\partial V}{\partial r}$. Similarly, the torsional linear response can be expressed as

$$\Delta\varphi = -\left(H^{(\varphi)}\right)^{-1} \frac{\partial V}{\partial \varphi} = -\left(H^{(\varphi)}\right)^{-1} J^t \frac{\partial V}{\partial r}. \tag{14}$$

Given a force $-\partial V/\partial r$, torsional linear response allows to compute the perturbed torsion angles, and to construct the perturbed structure that conserves both lengths and bond angles.

3.6. Null model of a generic perturbation

The analytic framework of normal mode analysis allows us to develop a null model of the most likely conformation change in response to a generic perturbation. Our aim is to compare observed conformation changes with this null model, in order to identify conformation changes correlated with the intrinsic motion of the protein significantly more than expected by chance.

The contribution of normal mode α to the conformation change is proportional to their squared overlap, $c_\alpha^2 \equiv |(\Delta r, Mx^\alpha)|^2$. In order to make the computations more robust, we use a sigmoidal function to set $c_\alpha \approx 0$ for modes whose contribution to the conformation change is of the order of 0.2 Å, which is representative of the experimental error in crystal structures, and we do not consider low collectivity normal modes that effectively move fewer than 4 degrees of freedom, $\kappa[\{p_i\}] < 4$.

We hypothesize that, within the null model, on the average a normal mode contributes to a generic conformation change as much as it contributes to the thermal motion, so that our null model is

$$\tilde{c}_\alpha^2 \propto \omega_\alpha^{-2}. \tag{15}$$

We generalize the above equation defining a family of null models $\tilde{c}_{q,\alpha}^2 \propto \omega_\alpha^{-q}$ and determine the parameter q that best fits observed conformation changes. We will show in the Results section that the best parameter is close to $q = 2$, which is the value predicted by our simple hypothesis. Therefore, we adopt Eq. (15) as the null model of conformation change throughout this paper.

As proposed in Ref. [14], we measure the deviations from the null model through the following Pearson correlation coefficient ρ:

$$\rho = \mathrm{Corr}\left(c_\alpha^2 \omega_\alpha^2, \omega_\alpha^{-2}\right). \tag{16}$$

For a general null model with parameter q, we define ρ as the correlation coefficient between $c_\alpha^2 / \tilde{c}_{q,\alpha}^2 = c_\alpha^2 \omega_\alpha^q$ and ω_α^{-2}. If the null model is approximately correct, we expect $\rho \approx 0$, whereas $\rho > 0$ means that low frequency normal modes contribute to the conformation change more than expected by chance. This implies that the harmonic energy barrier Eq. (13) is smaller than expected under neutral response. On the other hand, if $\rho < 0$, low frequency normal modes contribute less than expected and we expect that the harmonic energy barrier tends to be larger.

3.7. B factors and scale of normal modes

To fix the energy and length scale of the model, we predict the thermal fluctuations of each atom through the TNM as $\langle r_i^2 \rangle = k_B T \sum_\alpha \omega_\alpha^{-2} \left|\vec{x}_i^{\,\alpha}\right|^2$,

and we fit them to the thermal fluctuations measured as crystallographic B factors. For NMR structures, for which the B factors are not available, we set the root mean square fluctuations equal to 0.5 Å.

4. Results

4.1. Energy barriers in the NMR ensemble of ubiquitin

We analyzed the conformation changes between the unbound X-ray structure of ubiquitin (1ubi) and its NMR ensemble (2k39), and within the NMR ensemble itself. NMR structures are supposed to represent the thermal ensemble of the protein. We assume that the X-ray structure represents the equilibrium structure. Consistently, we find that the effective contact energy of the crystal structure, computed with the contact energy parameters in Ref. [36], is lower than the effective contact energy of 84% of the NMR structures. We compute the harmonic energy barrier of the transition from the X-ray structure to NMR structures. This is mainly determined by the RMSD and by the projection on normal modes. Fig. 1A shows that $DeltaE/\left(\mathrm{RMSD}^2\right.$ (full circles) is strongly correlated with the parameter ρ, as we predicted in the previous section: conformation changes with large ρ project along low frequency normal modes more than expected in the null model, and can undergo large deformations with a comparably small harmonic energy barrier. The correlation coefficient is $r=0.73$.

We expect that the harmonic "energy barriers" between the crystal structure and NMR structures are of order k_BT, otherwise they would be very unlikely in the equilibrium ensemble. This is in fact true for most NMR structures, but some present higher energy. We assume that this is due to violations of the harmonic approximation. We define the energy barrier from X to A passing through an intermediate structure B as $\Delta E_1(X,A)=\min_B \Delta E_0(X,B)+\Delta E_1(B,A)$. Within the harmonic approximation, the energy cost should be the same for a direct transition or going through an intermediate structure, $\Delta E_1(X,A)=\Delta E_0(X,A)$. Fig. 1A shows that this is the case for some structures, whose harmonic energy barrier remains the same if we consider an intermediate step (empty squares), but not for all structures, in particular structures with large barriers and small ρ violate the harmonic approximation. Nevertheless, the energy barrier through an intermediate is still significantly correlated with the parameter ρ: correlation coefficient $r=-0.44$, Student's-$t=-5.2$. We show in Fig. 1B the harmonic energy barrier through an intermediate versus the difference of contact energy between the final NMR structure and the starting X-ray structure. It is interesting that the two energies are significantly correlated ($r=0.30$, $t=3.35$), despite the fact that energy barriers are obtained with a Go-like model that does not know anything about the protein sequence and the contact energies depend on the specific pairs of amino acids in contact. The figure also shows that the barriers through an intermediate are small, so that these structures are likely to happen in the equilibrium ensemble.

4.2. Null model of conformation changes

As explained in section 3.6, we consider the family of null models $\tilde{c}^2_{q,\alpha}\propto\omega_\alpha^{-q}$. We test these models on 6230 pairs of structures of the same monomeric protein with the same ligand present in different crystals. We expect that this data set is depleted of functional conformation changes and agrees better with the null model, although functional conformation changes are present in this set as well (see below).

We assess these null models by two different tests. First, we measure the Kullback–Leibler divergence between the observed distribution $P_\alpha=c^2_\alpha/\sum_{\alpha'}c^2_{\alpha'}$ and the null-model distribution $\tilde{P}_{q,\alpha}=\omega_\alpha^{-q}/\sum_{\alpha'}\omega_{\alpha'}^{-q}$, which is defined as $d_{KL}=\sum_\alpha P_\alpha\left(\log(P_\alpha)-\log\left(\tilde{P}_\alpha\right)\right)$. The minimum of the distance is found for q between 1.75 and 2.0, see Fig. 2A. Second, we divide P_α times the null model $\tilde{}P_{q,\alpha}$. The result is proportional to $c^2_\alpha\omega^q_\alpha$. We measure the correlation coefficient between this variable and ω_α^{-2}. If the null-model is unbiased, this correlation coefficient should be zero. The average correlation coefficient vanishes for q slightly larger than 2.0, see Fig. 2B.

Altogether, these data indicate that the null model with $q=2$ best fits the observed conformation changes. This null-model is also special in that it predicts that each normal mode contributes to the conformation change as much as it contributes to the thermal motion of the protein, i.e., it predicts that a generic perturbation resembles the thermal dynamics of the protein.

We show in Fig. 3A the distribution of the correlation coefficient between the contribution of a normal mode to the conformation change, c^2_α, and its contribution to the thermal fluctuations, ω_α^{-2}. These correlations are almost always significant, as expected from our null model. Their average value is approximately 0.5, which is larger than the average correlation between observed and predicted B factors, shown in Fig. 3B. This comparison is consistent with the fact that B factors are influenced by rigid body degrees of freedom and crystal contacts not taken into account in ENMs. These results

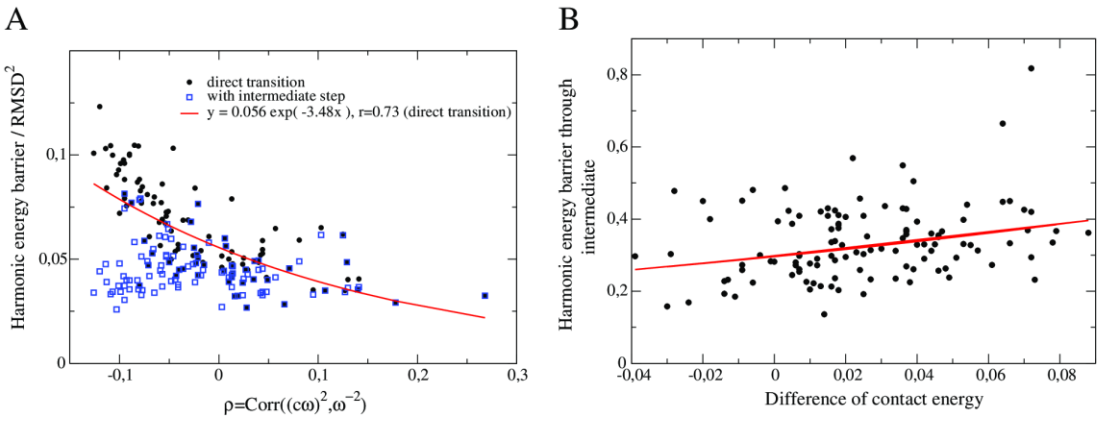


Fig. 1. A) Harmonic energy barrier between the unbound structure of ubiquitin studied by X-ray (PDB 1ubi) and its thermal ensemble studied by NMR (PDB 2k39). $\Delta E/\left(\mathrm{RMSD}^2\right)$ (circles) is negatively correlated with the parameter ρ, which is large if low frequency modes contribute more than expected to the conformation change. Empty squares represents energy barriers through an intermediate (see main text). B) Energy barriers through an intermediate versus the difference of the contact energy of the two structures.

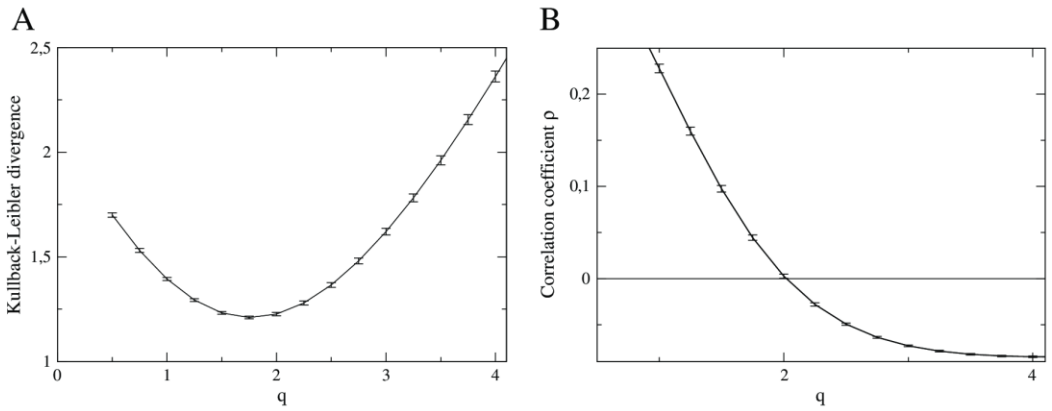


Fig. 2. A: Kullback–Leibler divergence between the observed distribution of the conformation change on normal modes, $P_\alpha = c_\alpha^2 / \sum_\alpha c_\alpha^2$, and the null-model distribution $\tilde{P}_{q,\alpha} = \omega_\alpha^{-q} / \sum_\alpha \omega_\alpha^{-q}$ as a function of q. B: Correlation coefficient between $c_\alpha^2 \omega_\alpha^q$ and ω_α^{-2}. The data set consists of 6230 pairs of monomeric structures with the same ligand. Here and in the following, error bars represent twice the standard error of the mean.

confirm that the null model described by Eq. (15) is a good starting point to model conformation changes.

4.3. Deviations from the null model

We measure deviations from the null-model through the parameter ρ, Eq. (16). We can see from Fig. 4A that the most likely values of ρ are close to zero, as predicted by the null model. However, we also observe large values of ρ that correspond to significant deviations from the null model.

The influence of these large values of ρ is clearly seen in Fig. 4B, and they are consistent with our theoretical expectation and with the results obtained for the NMR ensemble of ubiquitin: when ρ is large, harmonic energy barriers are small (not shown), the reciprocal collectivity of the normal mode is small, i.e., only few low-frequency normal modes significantly contribute to the conformation change (black curve), the maximum contribution of an individual mode to the conformation change is on the average larger than 20% (orange curve), the correlation between conformation change and thermal fluctuations is larger than 0.5 (green curve). Conversely, when the reciprocal collectivity is small, meaning that few modes contribute to the conformation change, ρ tends to be large, meaning that these modes are low frequency modes (see Fig. 4C), and when the harmonic energy barrier is small ρ (blue curve) and its significance, defined below (black curve) tend to be large, see Fig. 4D.

Since the parameter ρ is a correlation coefficient, its product times the square root of the number of normal modes used in the correlation should be distributed as a normal variable if there is no correlation. We therefore define the significance of ρ multiplying this variable times the square root of the number of normal modes. Its probability density is shown in Fig. 5A, where we can see that some conformation changes deviate very significantly from the null model. It is interesting to compare the different data sets. We find more significant pairs when the conformation change involves ligand binding (blue histograms) and homodimers (red) than monomeric proteins with the same ligands in both conformations (black).

The same ranking is found in different bins of the histogram, corresponding to independent pairs, which strongly suggest that the ranking is statistically significant.

Our results are consistent with an evolutionary interpretation. In evolutionary analysis, it is customary to compare the statistical properties of the observed data with a null model. If there is no significant difference, we cannot reject the hypothesis that the observed properties are due to chance. If the difference is significant, this suggests that natural selection may be responsible for the observed property. In the present case, we interpret a significant value of ρ as an indication of

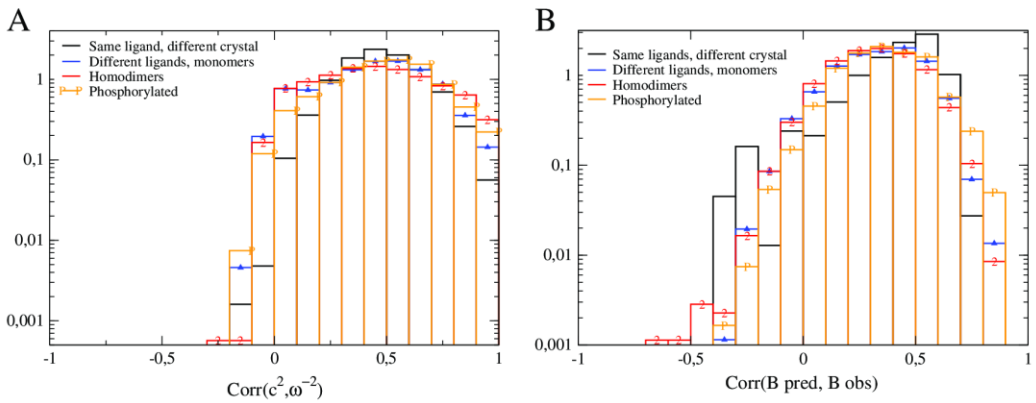


Fig. 3. For most proteins, the TNM yields highly significant correlations between the contribution of normal modes to conformation changes and to thermal dynamics (A), which are even larger than the correlations between predicted and observed B factors (B).

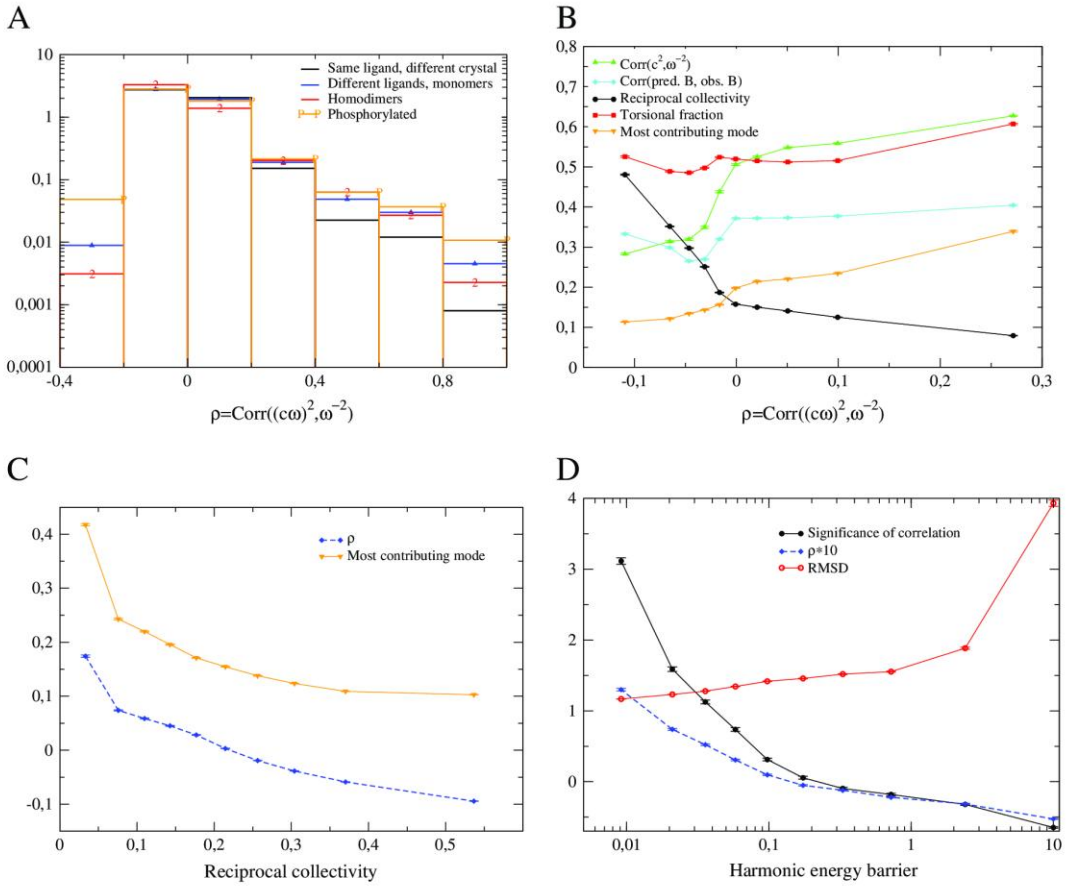


Fig. 4. A) Probability density of the parameter ρ, measuring deviations from the null model. B) Positive ρ implies that the reciprocal collectivity is small, i.e., fewer normal modes contribute to the conformation change. C) Conversely, for small reciprocal collectivity both ρ (blue) and the maximum contribution of an individual mode to the conformation change (orange) tend to be large. D) When the harmonic energy barriers are small ρ tends to be large, and its significance tends to be large (black curve).

natural selection. Since the effect of $\rho > 0$ is to reduce the free energy barrier of the conformation change, we interpret significant positive ρ as a hint that natural selection acted to reduce the energy barrier. This interpretation is consistent with the observation that this situation is more common in cases of ligand binding or homodimeric proteins than for monomeric proteins having the same ligands in the two conformations.

Likewise, we interpret significant and negative values of ρ as an indication that natural selection acted to increase the energy barrier to a conformation change. This may be favored by natural selection if the conformation changes should not happen spontaneously. However, the significance of negative values of ρ is small for most pairs. Interestingly, both significantly positive and negative values of ρ are more likely for phosphorylated/unphosphorylated pairs. In particular, significantly negative ρ is consistent with the view that the activation of certain regulatory proteins through phosphorylation should not happen spontaneously.

Fig. 5B shows qualitatively similar results for the probability density of the maximum contribution of a normal mode to the conformation change. Cases in which there is a mode that contributes more than 50% to the conformation change are more frequent for chains forming homodimers and for phosphorylated/unphosphorylated pairs, then for conformation changes involved in ligand binding and last for conformation changes without ligand binding. The ranking is also similar for conformation changes with small reciprocal collectivity, Fig. 5C, and for conformation changes with small harmonic energy barriers, Fig. 5D.

Note that the conformation change may have a functional meaning even when ligand binding is not involved. We examined pairs having highly significant values of ρ in the subset of monomeric proteins with the same ligand, and for most of them we found indications that they may represent functional conformation changes, for instance open and closed forms of an enzyme bound to a ligand (open and closed forms of maltodextrin-binding protein bound to maltose, PDB 1jw5 and 1anf), functional regulation by pH change (BACE1, an enzyme responsible for amyloid beta protein production, is active between pH 3.5 and 5.5, PDB 2zht and 2zhv, and inactive at pH 7, PDB 1zhv), sometimes in combination with calcium concentration (gelsolin at high calcium and pH = 7.5, PDB 1p8x, or low calcium and pH = 4.5, PDB 2fh1). In some cases, the two structures related by large ρ crystallize in different unit cells, they have almost identical folding stability as predicted by a contact energy function [36], the energy barrier is predicted to be small and one normal mode explains more than 50% of the conformation change, for instance human MST3 kinase, PDB 3a7f and 3a7g or chitinase from *Vibrio harveyi* complexed with an inhibitor, PDB 3ary and 3arz. This suggests that these pairs are almost equiprobable conformational substates that rapidly interconvert in solution. The fact that these conformation changes are described by the lowest frequency normal mode hints at their possible functional relevance, which should be addressed through experiments.

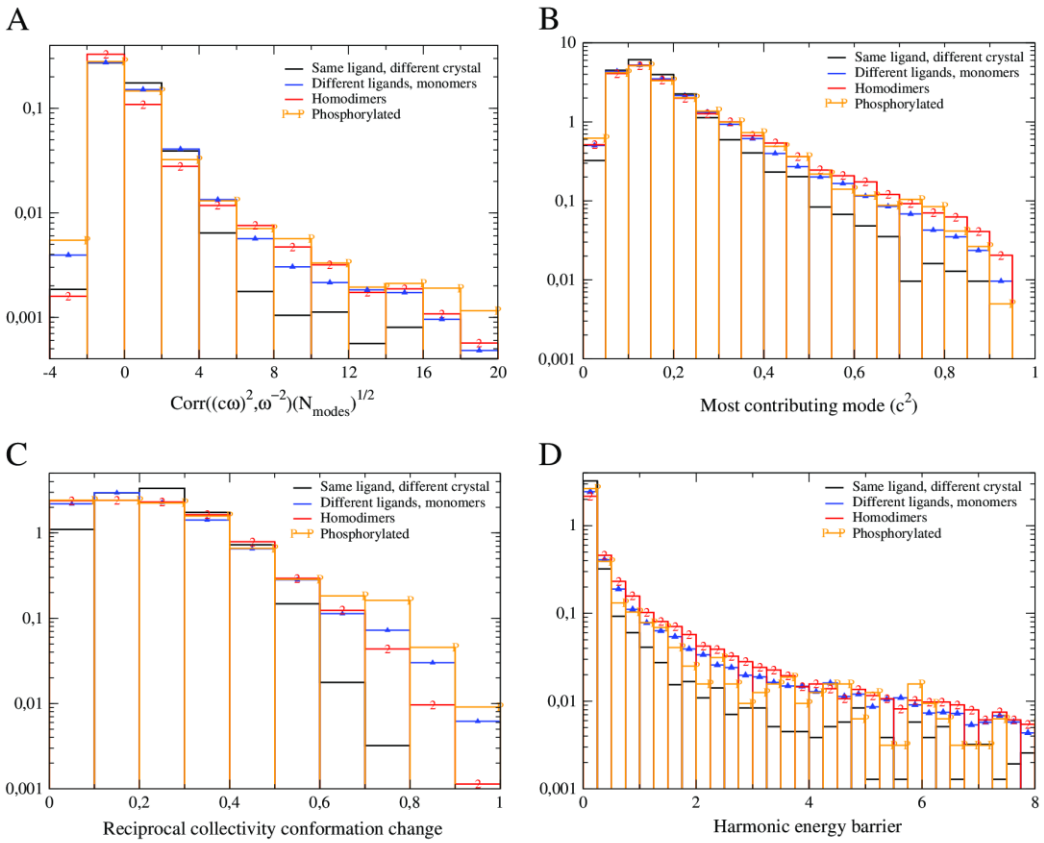


Fig. 5. Comparison of the probability densities of the significance of ρ (A), the maximum contribution of a normal mode to the conformation change (B), the fraction of normal modes that contribute to the conformation change (C) and the harmonic energy barriers (D) for different data sets.

4.4. Quality of the null model

We can see from Fig. 4A that slightly negative values of ρ are more frequent than slightly positive values. Fig. 4B sheds some light on this finding, since we can see that these slightly negative values of ρ are characterized by low values of the torsional fraction (red curve), meaning that many relevant degrees of freedom are not represented in the model, by low correlations between observed and predicted B factors, a hint that the TNM does not reproduce observed thermal fluctuations well, and by a small correlation between conformation change and thermal fluctuations, indicating that the null model is of poor quality. We interpret these findings in the sense that, when the quality of the model or of the data is low, the correlation between conformation change and thermal fluctuations and the parameter ρ are underestimated.

We present more results on the model quality in Fig. 6. A variable that strongly influences the quality of the results is the RMSD of the conformation change. When the RMSD is small, the conformation change is of a similar order as the experimental error and data are very noisy. Because of this reason, we exclude pairs with less than 1 Å RMSD from our analysis. At small RMSD, the correlation between conformation changes and thermal fluctuations is low, see Fig. 6A. We measured this correlation for pairs with RMSD smaller than 0.2 Å, finding that its average value is 0.30 ± 0.04, much smaller than 0.50 ± 0.06 for RMSD > 1 Å. Moreover, for small RMSD the torsional fraction is small, suggesting that deviations in bond angles and bond lengths due to the experimental error dominate the conformation change.

We illustrate in Fig. 6B the effect of the correlation between observed and predicted B factors, which is a measure of the quality of the TNM model. When this correlation is small, ρ is on the average negative (blue curve), the correlation between predicted thermal fluctuations and conformation change is very small (green curve), and many normal modes contribute to the conformation change (black curve).

Another relevant variable is the torsional fraction, whose distribution is bimodal, with a broad peak at a torsional fraction of 0.5 and a small peak close to 1, Fig. 6C. Torsional fractions smaller than 0.5 imply that ρ is on the average negative and the correlation between predicted thermal fluctuations and conformation change is very small, see blue and green curves in Fig. 6D.

Altogether, these results show that, if the quality of the model is small either because the conformation change is dominated by the experimental error on the coordinates (small torsional fraction and small RMSD), or because the TNM is a poor approximation of the thermal motion (badly predicted B factors), the value of ρ is systematically negative. However ρ is on the average very small even in these cases, so that the null model is not too bad.

4.5. Further direction: evolutionary conservation

In a number of interesting papers, Zheng and coworkers showed that evolutionary conservation, coupled with normal mode analysis, allows the identification of the network of dynamically important residues. These are residues that are evolutionarily conserved and are

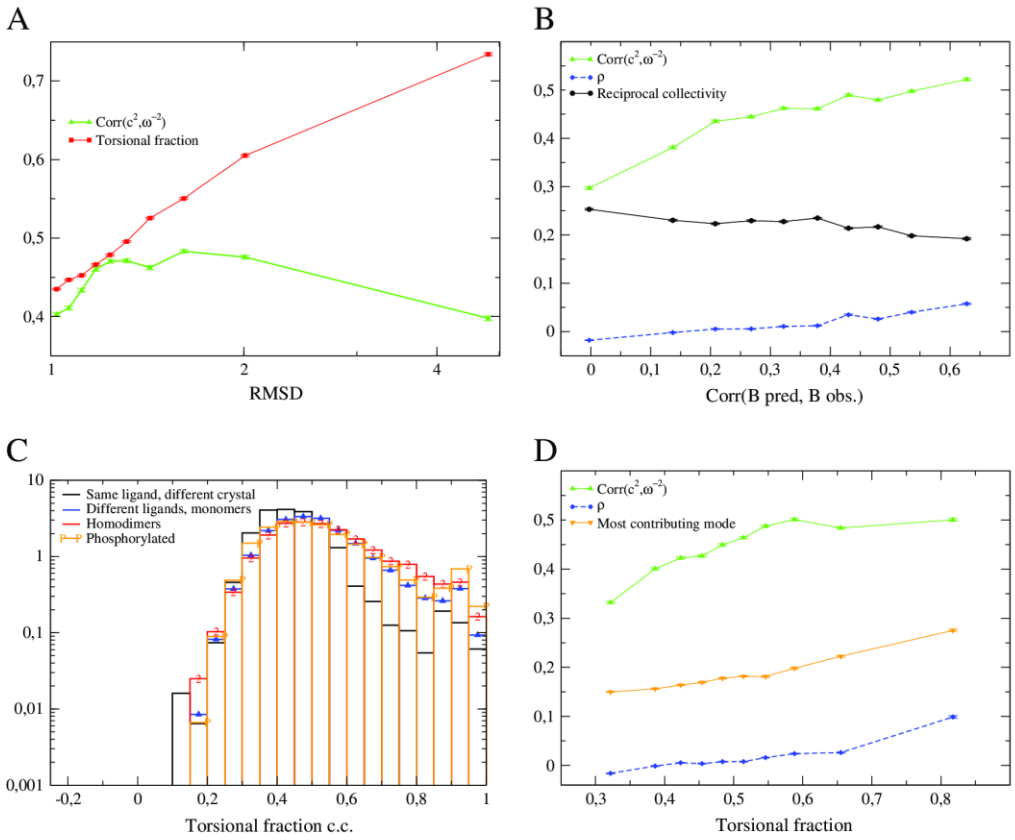


Fig. 6. Insufficient signal to noise ratio of the conformation change or low quality of the ENM or limit the quality of the null model. A) The correlation between conformation change and predicted thermal fluctuations $\mathrm{Corr}(c^2,\omega^{-2})$ is small both when the RMSD is small and large, and the torsional fraction increases with RMSD. B) $\mathrm{Corr}(c^2,\omega^{-2})$ and ρ are small when the correlation between observed and predicted B factors is small. C) Probability density of the torsional fraction. D) ρ and the correlation between predicted thermal fluctuations and conformation change are small for small torsional fraction.

strongly dynamically coupled in the normal mode that contributes most to the functional conformation change [37]. Moreover, they showed that one can identify the normal mode that contributes most to the functional dynamics as the normal mode whose dynamically coupled residues are more evolutionarily conserved [38].

The TNM allows the addressing of these issues with some technical improvements. On one hand, the normal modes that yield the largest contribution to the 6 functional conformation changes examined in Refs. [37,38] have on the average an overlap $|c_\alpha|/\sqrt{\sum_{alpha}|c_\alpha|^2}$ equal to 0.71 with the TNM, larger than the average overlap of 0.625 obtained in the original papers, see Table 1. Moreover, we developed two new dynamical couplings, the allosteric coupling and the co-directionality coupling (Bastolla, unpublished). The co-directionality coupling identifies pairs of residues whose direction of thermal motion are correlated, and the peaks derived from this coupling are closely related to those derived from the dynamical couplings of Ref. [37], without the need to specify which mode contributes most to the conformation change, as it is done in Ref. [37]. We plan to perform this analysis together with the conservation analysis proposed in Refs. [37,38], in order to identify pairs of residues that are dynamically coupled and experienced correlated mutations during evolution.

Table 1
Comparison of the analysis of six functional conformation changes with the TNM (third column) and with the ENM used in Refs. [37,38] (fourth column). The TNM and ENM columns report the mode that yields the largest overlap with the conformation change and the corresponding overlap.

Protein	PDB	TNM		ENM	
DNA pol. I, Taq	2ktq→3ktq	5	0.75	4	0.50
RNA pol. I, T7	1aro→1cez	1	0.84	1	0.66
DNA pol. β, human	1bpx→1bpy	1	0.72	1	0.71
RT, HIV-1	1n5y→1rtd	3	0.56	4	0.51
Myosin II, *D. discoideum*	1von→1mma	2	0.66	1	0.56
GroEL, *E. coli*	1aon→1grl	1	0.75	1	0.81

5. Discussion

In this work, we have shown that a simple computational analysis based on the torsional network model can give insight on conformation changes in proteins. We proposed a null-model of "random" (maximum entropy) conformation changes based on linear response theory. This model predicts that the most likely contribution of a normal mode to a conformation change is proportional to its contribution to thermal fluctuations, in such a way that low frequency normal modes give larger contributions to the conformation changes.

The agreement between the null model and observed conformation changes is good, however it is limited by two factors: (1) the quality of the TNM that we use to predict thermal fluctuations: if the agreement between observed and predicted thermal fluctuations is not good, then the null model is not well fulfilled; and (2) the signal to noise ratio: if the RMSD of the conformation change is small, then the experimental

error yields a large contribution to the conformation change and variations in bond angles and bond length, which are not considered in the TNM, give relevant contributions to the conformation change.

Deviations from the null model, measured by the parameter ρ, are generally small, but sometimes they are large and significant. These significant deviations indicate that the perturbation that produces the conformation change is not generic, but it is correlated with the intrinsic motion of the protein more than expected by chance. We attribute this situation as due to the result of natural selection, in the sense that the conformation change is functionally important and it produced a selective pressure on the evolution of the protein structure and intrinsic dynamics. A large value of ρ implies that only a small number of low frequency normal modes significantly contribute to the conformation change, and that the harmonic energy barrier opposed to the conformation change is small. Conformation changes with significant ρ are more frequently associated with ligand binding, in particular phosphorylation and formation of homodimers, than to conformation changes in which the ligands are the same, which in most cases correspond to different experimental representations of the same experimental system. However, even pairs having the same ligand may have a functional meaning, for instance when they represent the open and closed structures of an enzyme or regulation by pH or ionic strength.

The deep relationship between the thermal dynamics of a protein, represented by its normal modes, and its functional conformation changes brings a new perspective on the characterization of protein–ligand binding as conformational selection [39–41] or induced fit [42], reconciling the two models. Our model is based on linear response, therefore it represents induced fits, but it also shows that the functional conformation change pre-exists in the equilibrium dynamics of the protein. In order to distinguish between conformational selection and induced fit, energy barriers play a key role: If the barriers are small even in the absence of the ligand, the conformation change is likely to precede the binding, which can be described with the conformational selection model. Therefore, large values of ρ, which predict that the energy barriers are small, are likely associated with conformational selection, and they suggest that natural selection has achieved a protein whose intrinsic dynamics is intimately related with the desired functional dynamics.

Finally, the models presented here are relevant in the context of protein structure evolution. The computational analysis of protein superfamilies has showed that evolutionary variations of protein structures tend to happen more frequently in the subspace spanned by low frequency normal modes [43], a property that can be rationalized by modeling a mutation of the protein sequence as a perturbation whose influence on protein structure can be approximately predicted through linear response theory [44,45]. The model presented here goes one step forward, allowing the detection of evolutionary changes that are consistent with low frequency normal modes more than expected based on linear response theory.

Acknowledgements

We acknowledge insightful discussions with Alberto Pascual-García, and the financial support from the Comunidad de Madrid (Amarouto program to UB), the Spanish CSIC (JAE doc fellowship to RMG), and the Spanish Ministry of Science (Consolider grants CSD2006-00023 and BFU2011-24595).

References

[1] Nussinov R (2013), this special issue.
[2] K. Henzler-Wildman, D. Kern, Dynamic personalities of proteins, Nature 450 (2007) 964–972.
[3] N.M. Goodey, S.J. Benkovic, Allosteric regulation and catalysis emerge via a common route, Nat. Chem. Biol. 4 (2008) 474–482.
[4] L. Meireles, M. Gur, A. Bakan, I. Bahar, Pre-existing soft modes of motion uniquely defined by native contact topology facilitate ligand binding to proteins, Protein Sci. 20 (2011) 1645–1658.
[5] A. del Sol, C.J. Tsai, B. Ma, R. Nussinov, The origin of allosteric functional modulation: multiple pre-existing pathways, Structure 17 (2009) 1042–1050.
[6] M.M. Tirion, Large amplitude elastic motions in proteins from a single-parameter, atomic analysis, Phys. Rev. Lett. 77 (1996) 1905–1908.
[7] K. Hinsen, Analysis of domain motions by approximate normal mode calculations, Proteins 33 (1998) 417–429.
[8] A.R. Atilgan, S.R. Durell, R.L. Jernigan, M.C. Demirel, O. Keskin, I. Bahar, Anisotropy of fluctuation dynamics of proteins with an elastic network model, Biophys. J. 80 (2001) 505–515.
[9] H. Taketomi, Y. Ueda, N. Go, Studies on protein folding, unfolding and fluctuations by computer simulation. 1. The effect of specific amino acid sequence represented by specific inter-unit interactions, Int. J. Pept. Protein Res. 7 (1975) 44559.
[10] H.S. Chan, Z. Zhang, S. Wallin, Z. Liu, Cooperativity, local-nonlocal coupling, and nonnative interactions: principles of protein folding from coarse-grained models, Annu. Rev. Phys. Chem. 62 (2011) 301–326.
[11] J.D. Bryngelson, P.G. Wolynes, Spin glasses and the statistical mechanics of protein folding, Proc. Natl. Acad. Sci. U.S.A. 84 (1987) 7524–7528.
[12] C. Micheletti, P. Carloni, A. Maritan, Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models, Proteins 55 (2004) 635–645.
[13] I. Bahar, A.J. Rader, Coarse-grained normal mode analysis in structural biology, Curr. Opin. Struct. Biol. 15 (2005) 586–592.
[14] R. Mendez, U. Bastolla, Torsional network model: normal modes in torsion angle space better correlate with conformation changes in proteins, Phys. Rev. Lett. 104 (2010) 228103.
[15] N. Go, T. Noguti, T. Nishikawa, Dynamics of a small globular protein in terms of low-frequency vibrational modes, Proc. Natl. Acad. Sci. U.S.A. 80 (1983) 3696–3700.
[16] M. Lu, J. Ma, A minimalist network model for coarse-grained normal mode analysis and its application to biomolecular X-ray crystallography, Proc. Natl. Acad. Sci. U.S.A. 2008 (105) (2008) 15358–15363.
[17] J.R. Lopez-Blanco, J.I. Garzón, P. Chacón, iMod: multipurpose normal mode analysis in internal coordinates, Bioinformatics 15 (2011) 2843–2850.
[18] W. Nishima, G. Qi, S. Hayward, A. Kitao, DTA: dihedral transition analysis for characterization of the effects of large main-chain dihedral changes in proteins, Bioinformatics 25 (2009) 628–635.
[19] J.K. Bray, D.R. Weiss, M. Levitt, Optimized torsion-angle normal modes reproduce conformational changes more accurately than cartesian modes, Biophys. J. 101 (2011) 2966–2969.
[20] R. Goldstein, Classical Mechanics, Addison-Wesley, 1950.
[21] E. Eyal, C. Chennubhotla, L.W. Yang, I. Bahar, Anisotropic fluctuations of amino acids in protein structures: insights from X-ray crystallography and elastic network models, Bioinformatics 23 (2007) i175–i184.
[22] L. Yang, G. Song, R.L. Jernigan, Protein elastic network models and the ranges of cooperativity, Proc. Natl. Acad. Sci. U.S.A. 106 (2009) 12347–12352.
[23] M.F. Thorpe, Comment on elastic network models and proteins, Phys. Biol. 4 (2007) 60–63.
[24] M. Rueda, P. Chacón, M. Orozco, Thorough validation of protein normal mode analysis: a comparative study with essential dynamics, Structure 15 (2007) 565–575.
[25] D.A. Kondrashov, A.W. Van Wynsberghe, R.M. Bannen, Q. Cui, G.N. Phillips Jr., Protein structural variation in computational models and crystallographic data, Structure 15 (2007) 169–177.
[26] F. Tama, Y.H. Sanejouand, Conformational change of proteins arising from normal mode calculations, Protein Eng. 14 (2001) 1–6.
[27] D. Tobi, I. Bahar, Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state, Proc. Natl. Acad. Sci. U.S.A. 102 (2005) 18908–18913.
[28] S.E. Dobbins, V.I. Lesk, M.J. Sternberg, Insights into protein flexibility: The relationship between normal modes and conformational change upon protein–protein docking, Proc. Natl. Acad. Sci. U.S.A. 105 (2008) 10390–10395.
[29] W. Zheng, B.R. Brooks, D. Thirumalai, Allosteric transitions in biological nanomachines are described by robust normal modes of elastic networks, Curr. Protein Pept. Sci. 10 (2009) 128–132.
[30] M. Ikeguchi, J. Ueno, M. Sato, A. Kidera, Protein structural change upon ligand binding: linear response theory, Phys. Rev. Lett. 94 (2005) 078102.
[31] S. Omori, S. Fuchigami, M. Ikeguchi, A. Kidera, Linear response theory in dihedral angle space for protein structural change upon ligand binding, J. Comp. Chem. 30 (2009) 2602–2608.
[32] C. Eckart, Some studies concerning rotating axes and polyatomic molecules, Phys. Rev. 47 (1935) 552–558.
[33] B.H. Park, M. Levitt, The complexity and accuracy of discrete state models of protein structure, J. Mol. Biol. 249 (1995) 493–507.
[34] R. Bruschweiler, Collective protein dynamics and nuclear-spin relaxation, J. Chem. Phys. 102 (1995) 3396–3403.
[35] S.C. Flores, L.J. Lu, J. Yang, N. Carriero, M.B. Gerstein, Hinge Atlas: relating protein sequence to sites of structural flexibility, BMC Bioinforma. 8 (2007) 167.
[36] U. Bastolla, M. Vendruscolo, E.W. Knapp, A statistical mechanical method to optimize energy functions for protein folding, Proc. Natl. Acad. Sci. U.S.A. 97 (2000) 3977–3981.
[37] W. Zheng, B.R. Brooks, S. Doniach, D. Thirumalai, Network of dynamically important residues in the open/closed transition in polymerases is strongly conserved, Structure 13 (2005) 565–577.

[38] W. Zheng, B.R. Brooks, D. Thirumalai, Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations, Proc. Natl. Acad. Sci. U.S.A. 103 (2006) 7664–7669.
[39] J. Monod, J. Wyman, J.-P. Changeux, On the nature of allosteric transitions: a plausible model, J. Mol. Biol. 12 (1965) 88118.
[40] D.D. Boehr, R. Nussinov, P.E. Wright, The role of dynamic conformational ensembles in biomolecular recognition, Nat. Chem. Biol. 5 (2009) 78996.
[41] J.P. Changeux, S. Edelstein, Conformational selection or induced fit? 50 years of debate resolved F1000, Biol. Rep. 3 (2011) 19.
[42] D.E. Koshland Jr., Enzyme flexibility and enzyme action, J. Cell. Comp. Physiol. 54 (1959) 24558.
[43] A. Leo-Macias, P. Lopez-Romero, D. Lupyan, D. Zerbino, A.R. Ortiz, An analysis of core deformations in protein superfamilies, Biophys. J. 88 (2005) 1291–1299.
[44] J. Echave, Evolutionary divergence of protein structure: the linearly forced elastic network model, Chem. Phys. Lett. 457 (2008) 413–416.
[45] J. Echave, F.M. Fernández, A perturbative view of protein structural variation, Proteins 78 (2010) 173–180.

4.5.- Mejoras en la predicción teórica de interacciones moleculares para el diseño de fármacos asistido por ordenador

4.5.1.- Introducción y aportación del autor

Como se ha explicado en la sección 2.4.- Interacciones moleculares: *docking* y cribado virtual, podemos predecir el modo y la energía de unión entre una proteína y un potencial fármaco mediante técnicas de *docking* si disponemos de sus estructuras 3D. Nuestro laboratorio ha desarrollado un protocolo automático de búsqueda de nuevos fármacos basado en *docking* y VS detallado en la sección 2.4.3. En el trabajo que se presenta a continuación, nuestra plataforma ha sido mejorada en dos componentes esenciales: **(1)** el módulo que estima la energía de interacción entre proteína y ligando, en particular con respecto a los enlaces de hidrógeno, cuya contribución es esencial para la estabilidad del complejo y la especificidad de la interacción(artículo 5) y **(2)** el módulo que refina un conjunto representativo de conformaciones del ligando, explorando su flexibilidad, que tiene un papel esencial en el proceso de unión (artículo 6).

La contribución de la autora de esta tesis se ha centrado en los siguientes puntos principales: **(1)** la implementación de una función para determinar y cuantificar la contribución de los enlaces de hidrógeno (HB) a la unión proteína-ligando basándonos en parámetros geométricos, **(2)** la implementación de un algoritmo para la minimización de la energía de interacción proteína-ligando mediante la optimización de los ángulos torsionales del ligando en el contexto del sitio activo durante la fase de refinado de las soluciones de *docking* con el fin de mejorar la complementariedad entre el ligando y el receptor, **(3)** La integración y evaluación de estas nuevas funcionalidades en el protocolo de *docking* y cribado virtual desarrollado en nuestro laboratorio y **(4)** la búsqueda de nuevos fármacos para la diana terapéutica OXA-24 en colaboración con el grupo del Dr. Antonio Romero (Centro de Investigaciones Biológicas, CSIC).

En esta sección se presentarán dos artículos metodológicos y un proyecto de VS en curso.

4.5.2.- Incorporación del término de enlace de hidrógeno en la función de scoring MM-ISMSA

En el artículo (Klett et al., 2012) presentamos MM-ISMSA, una función de puntuación ultra-rápida y precisa que permite evaluar la energía de interacción de complejos moleculares, aplicable tanto a la evaluación de conformaciones obtenidas durante el muestreo del *docking*

como en el análisis de interacciones en trayectorias de MD. La función propuesta incluye **(1)** un término de mecánica molecular (MM) basado en el potencial 12-6 de Lennard-Jones, **(2)** una componente electrostática basada en el modelo ISM que permite el cálculo de las energías libres de desolvatación individuales de las proteínas y los ligandos que forman el complejo más un término para los HB y **(3)** la contribución por pérdida de área de superficie expuesta al solvente (SA) para tener en cuenta la disminución en el número de contactos con las moléculas de agua cuando se forma el complejo. El método MM-ISMSA, al igual que los módulos MM-GBSA y MM-PBSA de AMBER, permite la descomposición de la energía de unión por pares de residuos y por residuos individuales, favoreciendo una fácil identificación de aquellos que son responsables de la formación y estabilidad de los complejos. Para facilitar su manejo a usuarios no expertos, el programa se ha incluido como un *plugin* dentro del visualizador de estructuras tridimensionales PyMOL (http://www.pymol.org).

La aportación de la autora de esta tesis al proyecto ha sido la de incorporar un nuevo término energético de HB en la función de *scoring* MM-ISMSA. La puntuación de cada HB está basada en la función de puntuación ChemScore del programa GOLD (Verdonk, Cole, Hartshorn, Murray, & Taylor, 2003). ChemScore usa una función de bloque *ad hoc* (Figura 24) para describir términos de contacto de varios tipos en función de parámetros geométricos como distancias y ángulos, en la cual se distinguen un rango ideal donde la puntuación es máxima, un rango donde la puntuación es cero y una interpolación lineal entre ambos. La puntuación de un HB depende de las variables geométricas de distancia, el ángulo α (*i.e.* ángulo comprendido entre los átomos donador HB, hidrógeno y aceptor de HB) y el ángulo β (*i.e.* ángulo comprendido entre los átomos hidrógeno, aceptor de HB y átomo directamente unido a aceptor de HB) y del tipo de átomos involucrados en la interacción.

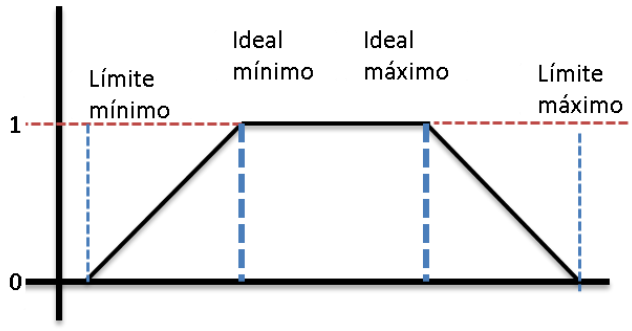


Figura 24. Función de bloque basada en ChemScore (GOLD) para evaluar la bondad de la geometría de un HB (variable x). La contribución energética estimada surge de multiplicar la puntuación [Score(x)] por un parámetro dependiente del tipo de átomos involucrados en el HB.

Para la obtención de los parámetros se llevó a cabo el análisis estadístico de las variables geométricas definidas para los enlaces de hidrógeno presentes en un conjunto de 85 complejos cristalográficos proteína-ligando obtenidos a partir de su inspección visual e información bibliográfica [Conjunto diverso de Astex (Hartshorn et al., 2007), conjunto de entrenamiento]. Durante la validación del nuevo programa usamos un conjunto de prueba de 23 proteínas incluido en el artículo previo de ISM (Morreale et al., 2007). Para discriminar contactos polares de HB, el átomo de hidrógeno fue considerado en las medidas. Obtuvimos parámetros de rangos ideales y rangos límites para cada variable geométrica considerada (Tabla 1) y observamos que sus distribuciones no presentaban simetría en el caso del ángulo α y la distancia, pero sí para el ángulo α (Figura 25).

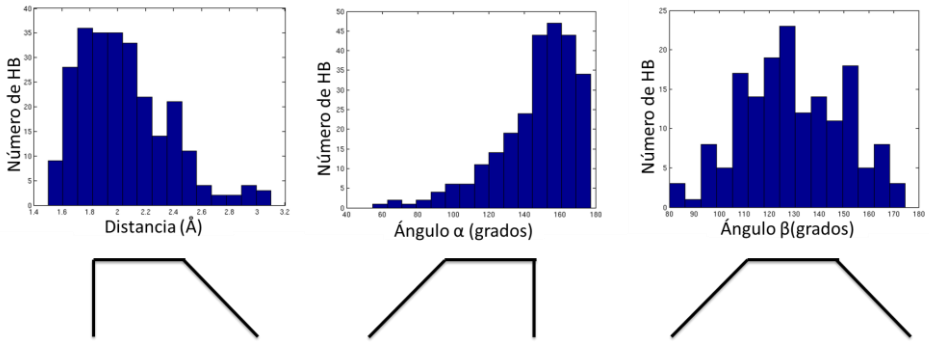


Figura 25. Distribuciones de las variables geométricas medidas: distancia, ángulo α y ángulo β. Debajo de cada una de ellas esquematizamos la forma de la distribución de datos obtenida.

	Límite mínimo	Mínimo ideal	Máximo ideal	Límite máximo
Distancia (Å)	1,5	1,8	2,4	2,7
Ángulo α (grados)	100	130	165	180
Ángulo β (grados)	90	115	145	180

Tabla 1. Valores de los rangos preferentes para las variables geométricas de los HB en base al análisis estadístico: distancia, ángulo α y ángulo β. Parámetros obtenidos a partir del conjunto diverso de Astex (85 complejos proteína-ligando). Dado que la forma de la distribuciones de los datos en el caso de la distancia y el ángulo α no son simétricas (Figura 25), alguno de estos parámetros fueron modificados *a posteriori*.

Los resultados mostraron que el 92,5% de los HB del conjunto de entrenamiento y el 83,7% de los HB del conjunto de prueba fueron correctamente identificados por el programa, en comparación con el método previo que recuperaba tan sólo el 54% de los HB del conjunto de Astex. La nueva implementación dentro del programa de *docking* repercute en mejores resultados en el 25% de los casos del conjunto de Astex. Tras la nueva parametrización,

obtenemos una identificación y cuantificación de los HB menos restrictiva a nivel geométrico y más precisa a nivel de su puntuación asociada, lo que favorece una mayor fiabilidad en la predicción de la unión proteína-ligando durante el *docking*. En resumen, presentamos una función de *scoring* (MM-ISMSA) precisa (80% de aciertos) y muy rápida (unos 5 segundos para sistemas de 1000 residuos), de gran utilidad durante la evaluación energética de complejos proteína-ligando y proteína-proteína en protocolos de *docking* y MD.

Artículo 5

JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

MM-ISMSA: An Ultrafast and Accurate Scoring Function for Protein−Protein Docking

Javier Klett,[†] Alfonso Núñez-Salgado,[†] Helena G. Dos Santos,[†] Álvaro Cortés-Cabrera,[†,‡] Almudena Perona,[†,§] Rubén Gil-Redondo,[†,§] David Abia,[†] Federico Gago,[‡] and Antonio Morreale*[,†]

[†]Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa (CSIC-UAM), Campus de Cantoblanco UAM, E-28049 Madrid, Spain

[‡]Departamento de Farmacología, Universidad de Alcalá, Alcalá de Henares, E-28871 Madrid, Spain

[§]SmartLigs Bioinformática S.L., Fundación Parque Científico de Madrid, c/Faraday, 7. Campus de Cantoblanco UAM, E-28049 Madrid, Spain

ABSTRACT: An ultrafast and accurate scoring function for protein−protein docking is presented. It includes (1) a molecular mechanics (MM) part based on a 12−6 Lennard-Jones potential; (2) an electrostatic component based on an implicit solvent model (ISM) with individual desolvation penalties for each partner in the protein−protein complex plus a hydrogen bonding term; and (3) a surface area (SA) contribution to account for the loss of water contacts upon protein−protein complex formation. The accuracy and performance of the scoring function, termed MM-ISMSA, have been assessed by (1) comparing the total binding energies, the electrostatic term, and its components (charge−charge and individual desolvation energies), as well as the per residue contributions, to results obtained with well-established methods such as APBSA or MM-PB(GB)SA for a set of 1242 decoy protein−protein complexes and (2) testing its ability to recognize the docking solution closest to the experimental structure as that providing the most favorable total binding energy. For this purpose, a test set consisting of 15 protein−protein complexes with known 3D structure mixed with 10 decoys for each complex was used. The correlation between the values afforded by MM-ISMSA and those from the other methods is quite remarkable ($r^2 \sim 0.9$), and only 0.2−5.0 s (depending on the number of residues) are spent on a single calculation including an *all vs all* pairwise energy decomposition. On the other hand, MM-ISMSA correctly identifies the best docking solution as that closest to the experimental structure in 80% of the cases. Finally, MM-ISMSA can process molecular dynamics trajectories and reports the results as averaged values with their standard deviations. MM-ISMSA has been implemented as a plugin to the widely used molecular graphics program PyMOL, although it can also be executed in command-line mode. MM-ISMSA is distributed free of charge to nonprofit organizations.

1. INTRODUCTION

Molecular association (binding) plays a key role in cellular function and communication, and many illnesses can be directly linked to an improper balance of interactions among distinct molecular species. Of special importance are those established between different proteins or between small molecules and proteins. In the latter case, we usually talk about ligands and receptors, but in the following, we will use this terminology to refer to the two binding partners. Understanding how binding takes place and how this event can be theoretically modeled is of paramount importance in today's drug discovery campaigns,[1] especially in fields like protein engineering,[2] ligand and fragment docking,[3] virtual screening (VS),[4] and computational mutagenesis,[5] among others.

A large number of methodological advances have been introduced since the first theoretical simulation of a biologically relevant system[6] that, in favorable cases, allow one to reproduce experimental binding affinities with an error comparable to that of the experimental measurements.[7] On top of that, modern computers, sometimes including tailor-made architectures (e.g., the Anton machine[8]), and supercomputers[9] allow researchers to undertake calculations that were unimaginable just a few years ago to address difficult problems such as simulating protein folding,[10] long MD of very large systems,[11] or unbiased drug binding patterns.[12]

Nevertheless, although theoretical methods and computer technologies are continuously improving, there are still some bottlenecks. Representative examples are intrinsically massive calculations, as undertaken in VS, where the number of molecules to be simulated can reach the order of several millions, and the analysis and interpretation of long MD trajectories especially when solvent effects must be properly accounted for, per-residue analyses must be performed, or one wishes to estimate elusive entropic effects.

Here, we are interested in the electrostatic contribution to the free energy of binding, and more specifically in the effect played by the solvent. There are two opposite, although complementary, ways to account for this effect:[13] either representing the solvent explicitly by means of a set of discrete water molecules surrounding the solute or *via* a mathematical function able to describe the behavior of the bulk solvent. There are also examples in which both methods have been combined in a sort of hybrid solvent model.[14] However, both have advantages and caveats, and selecting the most appropriate description for a particular study greatly depends on the computational power available, as explicit models require the calculation of a large amount of interactions due to the presence of numerous water molecules.

Received: June 15, 2012
Published: August 2, 2012

dx.doi.org/10.1021/ct300497z | *J. Chem. Theory Comput.* 2012, 8, 3395−3408

For docking-related tools, implicit models are usually preferred, as they are faster than their explicit counterpart while performing quite similarly. Implicit models start by solving the classical Poisson equation (PE).[15] But, in many cases it is too computationally expensive, and other alternatives such as the generalized Born (GB) model are employed instead.[16] In addition, solvation models based on group contributions (effective energy function, EFF1) have also been proposed and successfully employed in proteins[17] and protein−ligand force fields.[18]

In this paper, we extend our previously developed GB-like solvent model, called ISM (Implicit Solvent Model),[19] to the protein−protein docking problem and compare its performance to well-established methods like APBS (PE solver) and MM-PB(GB)SA (as implemented in AmberTools[20]) using two different test sets. For the procedures to be strictly comparable, we first incorporated MM (Molecular Mechanics) and SA (Surface Area) terms to the APBS (MM-APBSSA) and ISM (MM-ISMSA) methods. Comparisons between the three different techniques were performed in order to assess their relative speeds and to validate the different values provided by our method for total binding free energies, individual components, and per residue energy decompositions. Finally, to extend the usability of MM-ISMSA within the scientific community, we have developed a graphical user interface (GUI) that allows its use *via* the popular PyMOL program.[21] This plugin, which can be used to analyze single structures or complete MD trajectories, can be downloaded following free registration from the CBM Bioinformatics Unit's web page (http://ub.cbm.uam.es).

2. THEORETICAL BACKGROUND

2.1. Statistical Thermodynamics of Binding: Interaction Terms. According to classical thermodynamics,[22] molecular association can be described as an equilibrium ([R] + [L] ⇌ [RL], where R and L represent receptor and ligand, respectively, and RL, the complex formed between them) governed by the association and dissociation rate constants. The ratio between them is the equilibrium binding constant, K, which is related to the free energy change taking place in the process ($\Delta G_{binding}$) by the well-known equation:

$$\Delta G_{binding} = \Delta G_{RL} - \Delta G_{R} - \Delta G_{L} = -R_{gas} T \ln K \quad (1)$$

where R_{gas} is the universal gas constant, T is the temperature in Kelvin, and ΔG_X represents the free energy corresponding to the complex (X = RL), receptor (X = R), and ligand (X = L).

The calculation of $\Delta G_{binding}$ would entail extremely lengthy simulations in which the ligand diffuses into the receptor's binding site, but this is hardly ever done. As useful alternatives, the ligand can be "grown" slowly both in the bulk solvent and inside the binding site to calculate free energy differences,[7] or $\Delta G_{binding}$ can be estimated as the difference between the free energies of bound and unbound states, as in linear interaction energy (LIE)[23] and MM-PB(GB)SA[24] "end-point" methods.

It has been customary to describe the binding process through the thermodynamic cycle in Figure 1, where ΔG_{int} refers to the binding process in the gas phase, and ΔG^{RL}_{solv}, ΔG^{R}_{solv}, and ΔG^{L}_{solv} are the free energies of solvation for the complex, receptor, and ligand, respectively. Next, because $\Delta G_{binding}$ is a state variable, the cycle can be solved to yield:

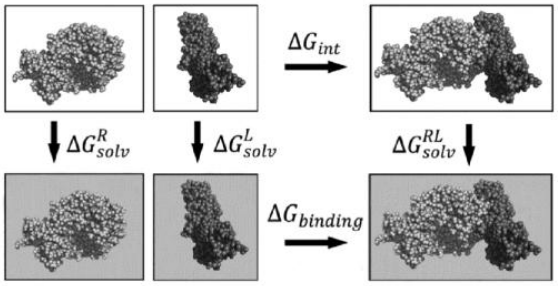


Figure 1. Graphical representation of the commonly used thermodynamic cycle to estimate $\Delta G_{binding}$. The shadowed boxes represent the systems (receptor, ligand, and complex) immersed in the solvent.

$$\Delta G_{binding} = \Delta G_{int} + \Delta G^{RL}_{solv} - \Delta G^{R}_{solv} - \Delta G^{L}_{solv} = \Delta G_{int} + \Delta G_{solv} \quad (2)$$

Usually, from Figure 1 and eq 2 the solvation contribution (ΔG_{solv}) is obtained as a single term (e.g., in GB) and, as a consequence, individual solvation energies for the ligand and receptor are not available. As an alternative, a different description of the binding process[25] can be considered (Figure 2) that consists of first desolvating the apposing surfaces of both ligand and receptor and then letting the charges of the two molecules interact.

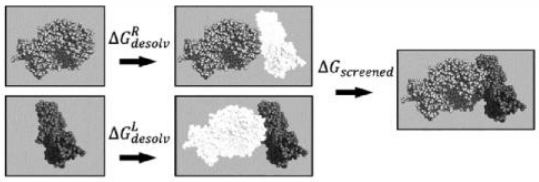


Figure 2. Alternative description of the binding process to estimate $\Delta G_{binding}$. This type of equilibrium is commonly used in combination with PE solvers. White molecules are uncharged, and they are used to replace the high-dielectric solvent with a low-dielectric medium in the desolvation calculations.

Thermodynamically, $\Delta G_{binding}$ has an enthalpic ($\Delta H_{binding}$) and an entropic ($\Delta S_{binding}$) component:

$$\Delta G_{binding} = \Delta H_{binding} - T\Delta S_{binding} \quad (3)$$

$\Delta H_{binding}$ contains van der Waals ($\Delta G^{vdW}_{binding}$), hydrogen bonding ($\Delta G^{hb}_{binding}$), and solvent-related contributions that can be further subdivided into polar ($\Delta G^{p}_{binding}$) and apolar ($\Delta G^{np}_{binding}$) components. The former contains the Coulombic interactions ($\Delta G^{elec,coul}_{binding}$) together with ligand ($\Delta G^{elec,desolv_R}_{binding}$) and receptor ($\Delta G^{elec,desolv_L}_{binding}$) desolvation terms, whereas the latter includes the cavitation term (the work required to create a cavity within the solvent to introduce the solute) and the van der Waals solute−solvent interactions. A linear relationship is usually assumed between the composite of the latter two components and the change in solvent-accessible surface area (SASA) of the ligand and receptor upon binding. On the other hand, the entropic contribution arises from the loss of some protein degrees of freedom that become frozen when the complex is formed and also from solvent reorganization, as some water molecules present within the binding site will be released to the bulk solvent

as a consequence of the binding event. This entropic contribution is rarely taken into account when $\Delta G_{binding}$ is computed due to its complexity, high computational demand, and slow convergence.[26] The electrostatic component is the most challenging term, and this will be the focus of the present work.

2.2. The PE Model. The classical way to deal with the electrostatic contribution to the binding energy (ΔG_{elec}) is by solving the PE, which relates the electrostatic potential $\phi(r)$ and the charge distribution $\rho(r)$:

$$\nabla[\varepsilon(r)\cdot\nabla\phi(r)] = -4\pi\rho(r) \tag{4}$$

where $\varepsilon(r)$ is a distance-dependent dielectric function. For a given $\rho(r)$, $\phi(r)$ can be calculated *via* the PE so that

$$\Delta G_{elec} = \frac{1}{2}\int \rho(r)\,\phi(r)\,\mathrm{d}v \tag{5}$$

Because the analytical solution of PE is possible only for very simple geometries, for biological molecules we have to rely on numerical methods such as finite differences,[27] finite elements,[28] or boundary elements.[29] Nonetheless, solving the PE is still a computationally demanding task in many molecular modeling areas, despite constant improvements over the years.[30]

Equation 5 can be used in different ways to obtain an estimation of the binding free energy, either by computing the desolvation terms of the thermodynamic cycle depicted in Figure 1 (as implemented in the MM-PBSA method) or by applying the cycle shown in Figure 2 as described in section 3.2.1, where the Coulombic contribution is obtained by computing the product of ligand charges times the electrostatic potential generated by the protein on the ligand charge centers. On the other hand, receptor and ligand electrostatic desolvation energies are calculated in two successive steps (Figure 2): a first one, where a calculation is performed for the receptor and ligand alone, and a second one, for the ligand in the complex, with uncharged receptor, and for the receptor in the complex, with uncharged ligand.

2.3. The GB Model. The GB model is based on the Born approximation and can be easily derived from the PE, assuming a spherical solute that has the whole charge located at its center, according to

$$\Delta G^{elec}_{solv}(\mathrm{Born}) = -166\left(1-\frac{1}{\varepsilon}\right)\frac{q^2}{r} \tag{6}$$

where ε is the dielectric constant of the solvent, and q and r are the charge and the radius of the sphere, respectively. Taking into account that molecules can be represented as a set of interacting spheres, eq 6 can be generalized to the expression commonly used in the GB models:

$$\begin{aligned}\Delta G^{elec}_{solv}(\mathrm{GB}) &= \Delta G_{vac} + \Delta G_{pol}\\ &= 332\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{q_i q_j}{r_{ij}} - 166\left(1-\frac{1}{\varepsilon}\right)\\ &\quad \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{f_{GB}}\end{aligned} \tag{7}$$

where N is the total number of atoms, r_{ij} is the distance between atoms i and j, q_i and q_j are the atomic charges of atoms i and j, and f_{GB} is the GB function defined as

$$f_{GB}(r_{ij}, i, j) = [r_{ij}^2 + \alpha_i\alpha_j\, e^{(-r_{ij}^2/4\alpha_i\alpha_j)}]^{1/2} \tag{8}$$

α is the so-called effective Born radius, which is the distance from an atom to the molecular surface. Note that f_{GB} is different depending on the system, that is, the receptor, the ligand, or the complex. In eq 7, the first term (ΔG_{vac}) represents the electrostatic interaction in vacuo, while the second (ΔG_{pol}) accounts for the *polarization* effects due to the solvent. In fact, it is this second term that is calculated as the electrostatic component of the free energy of solvation in GB-based methods, and as such it has been implemented in many different programs and, in particular, in the AmberTools package.

Applying eq 7 to the thermodynamic equilibrium depicted in Figure 1, it is possible to separate the electrostatic interaction between charges (in a vacuum and in the solvent) from the pure desolvation terms for the receptor and ligand:

$$\begin{aligned}\Delta\Delta G^{elec}_{solv}(\mathrm{GB}) &= 332\sum_{i=1}^{N_R}\sum_{j=1}^{N_L}\frac{q_i^R q_j^L}{r_{ij}} - 332\left(1-\frac{1}{\varepsilon}\right)\\ &\quad \sum_{i=1}^{N_R}\sum_{j=1}^{N_L}\frac{q_i^R q_j^L}{f^{RL}_{GB}} - 166\left(1-\frac{1}{\varepsilon}\right)\\ &\quad \sum_{i=1}^{N_R}\sum_{j=1}^{N_L}\left[\frac{q_i^R q_j^R}{f^{RL}_{GB}} - \frac{q_i^R q_j^R}{f^{R}_{GB}}\right]\\ &\quad - 166\left(1-\frac{1}{\varepsilon}\right)\sum_{i=1}^{N_R}\sum_{j=1}^{N_L}\left[\frac{q_i^L q_j^L}{f^{RL}_{GB}} - \frac{q_i^L q_j^L}{f^{L}_{GB}}\right]\end{aligned} \tag{9}$$

Unfortunately, as commented upon above, the GB method as implemented in AmberTools does not allow for such decomposition, and only the total polarization energy is obtained (ΔG_{pol} in eq 7). Nevertheless, setting the atomic charges in the receptor/ligand to zero and subtracting the resulting GB term from the GB term of a standard calculation would afford the ligand/receptor desolvation terms independently. Accordingly, zeroing the ligand's charges:

$$\begin{aligned}\Delta\Delta G^{R}_{desolv} &= \Delta G^{MMGBSA}_{elec}(\mathrm{GB}) - \Delta G^{MMGBSA}_{elec}(\mathrm{GB}, q_i^L = 0)\\ &= -166\left(1-\frac{1}{\varepsilon}\right)\sum_{i=1}^{N_R}\sum_{j=1}^{N_R}(q_i^R)^2\left[\frac{1}{f^{RL}_{GB}} - \frac{1}{f^{R}_{GB}}\right]\end{aligned} \tag{10}$$

The same can be done if the receptor's charges are zeroed:

$$\begin{aligned}\Delta\Delta G^{L}_{desolv} &= \Delta G^{MMGBSA}_{elec}(\mathrm{GB}) - \Delta G^{MMGBSA}_{elec}(\mathrm{GB}, q_i^R = 0)\\ &= -166\left(1-\frac{1}{\varepsilon}\right)\sum_{i=1}^{N_L}\sum_{j=1}^{N_L}(q_i^L)^2\left[\frac{1}{f^{RL}_{GB}} - \frac{1}{f^{L}_{GB}}\right]\end{aligned} \tag{11}$$

Choosing this alternative would require (a) manipulating the topology (*top*) files used by AmberTools to set the charges of the ligand (or the receptor) to zero and (b) an additional GB calculation (ΔG_{pol} in eq 7) to obtain the Coulombic interaction screened by the solvent ($\Delta G_{screened}$, the first right-hand term in eq 9). That is, taking into account

$$\Delta G_{pol} = \Delta G_{screened} + \Delta G^{R}_{desolv} + \Delta G^{L}_{desolv}$$

and the fact that we can calculate both desolvation energies as shown above:

$$\Delta G_{\text{screened}} = \Delta G^{\text{elec}}_{\text{solv}}(\text{GB}) - \Delta G^{\text{R}}_{\text{desolv}} - \Delta G^{\text{L}}_{\text{desolv}}$$

three calculations are needed to obtain the total free energy of solvation and its components.

2.4. The ISM Model. The model starts from the Lorentz–Debye–Sack theory of polar liquids,[31] which establishes that the screening effect due to the solvent shows a sigmoidal distance-dependent dielectric function of the form:

$$D(r) = \frac{\varepsilon + 1}{1 + k\, e^{-\lambda(\varepsilon+1)r}} - 1 \tag{12}$$

where ε is the solvent dielectric constant, $k = (\varepsilon - 1)/2$, λ is a parameter controlling the rate of change of $D(r)$, and r is the distance. ISM considers that the main contribution to the electrostatic desolvation of an atom originates from the displacement of the first shell of water molecules that surrounds that atom. Taking these two facts into account, ISM's starting equation, as proposed by Hassan et al.,[32] is the following:

$$\Delta G^{\text{elec}}_{\text{solv}}(\text{ISM}) = \sum_{i<j}^{N} \frac{q_i q_j}{r_{ij}} \left[\frac{1}{D_s(r_{ij})} - \frac{1}{D_v(r_{ij})} \right] + \frac{1}{2} \sum_{i}^{N} q_i^2 \left\{ \frac{1}{R_{i,\text{Bs}}} \left[\frac{1}{D_s(R_{i,\text{Bs}})} - 1 \right] - \frac{1}{R_{i,\text{Bv}}} \left[\frac{1}{D_s(R_{i,\text{Bv}})} - 1 \right] \right\} \tag{13}$$

where $R_{i,\text{Bs}}$ and $R_{i,\text{Bv}}$ are the effective Born radii for the processes of transferring an atom from a vacuum into a protein interior, surrounded by either solvent or a vacuum, respectively. The model has proven to be useful to study the structure and dynamics of proteins[33] and has been implemented as a solvation method within the molecular dynamics code of the program CHARMm.[34]

This model has been extended by us to deal with ligand–receptor[19] and protein–protein interactions (this work). Considering the thermodynamic cycle of binding depicted in Figure 1 and the expression for $\Delta G^{\text{elec}}_{\text{solv}}(\text{ISM})$:

$$\Delta G^{\text{ISM}}_{\text{elec}} = \frac{1}{2} \sum_{i=1}^{N_R} \sum_{j=1}^{N_L} \frac{q_i^{\text{R}} q_j^{\text{L}}}{r_{ij}} \left[\frac{1}{D_s(r_{ij})} \right] + \frac{1}{2} \sum_{i}^{N_{RL}} (q_i^{\text{RL}})^2 \left[\left(\frac{1}{D_s(R^{\text{c}}_{i,B_s}) R^{\text{c}}_{i,B_s}} - \frac{1}{D_s(R^{\text{u}}_{i,B_s}) R^{\text{u}}_{i,B_s}} \right) + \left(\frac{1}{R^{\text{u}}_{i,B_s}} - \frac{1}{R^{\text{c}}_{i,B_s}} \right) \right] \tag{14}$$

where the superscripts c and u stand for complexed and uncomplexed forms of both the ligand and receptor. There is an evident resemblance between ISM (eq 14) and GB (eq 9) formulations: the first term describes the interaction established between the receptor and ligand screened by the dielectric function, and the second is the desolvation penalty, accounted for by the difference between complexed and uncomplexed partners in terms of their Born radii and dielectric function in the solvent.

As an additional advantage, ISM directly yields individual desolvation terms without the need to perform any extra calculations, as commented upon before for GB. In fact, to isolate the desolvation term for the receptor, only the summation of the atoms concerning the receptor must be considered in the GB-like term (second term on the right-hand side of eq 14), which is the same as setting to zero the ligand's charges (eq 15) or, for the ligand desolvation term, setting to zero the receptor's charges (eq 16):

$$\Delta G^{\text{R}}_{\text{desolv}} = \frac{1}{2} \sum_{i}^{N_R} (q_i^{\text{R}})^2 \left[\frac{1}{R^{\text{c}}_{i,B_s}} \left(\frac{1}{D_s(R^{\text{c}}_{i,B_s})} - 1 \right) - \frac{1}{R^{\text{u}}_{i,B_s}} \left(\frac{1}{D_s(R^{\text{u}}_{i,B_s})} - 1 \right) \right] \tag{15}$$

$$\Delta G^{\text{L}}_{\text{desolv}} = \frac{1}{2} \sum_{i}^{N_L} (q_i^{\text{L}})^2 \left[\frac{1}{R^{\text{c}}_{i,B_s}} \left(\frac{1}{D_s(R^{\text{c}}_{i,B_s})} - 1 \right) - \frac{1}{R^{\text{u}}_{i,B_s}} \left(\frac{1}{D_s(R^{\text{u}}_{i,B_s})} - 1 \right) \right] \tag{16}$$

2.5. Pairwise Decomposition of the Binding Energy. The value of the free energy of binding as a whole is useful when comparing the binding strength of a set of ligands toward a target of interest. However, it does not provide any information on the relative contributions of individual residues. Knowing which residues are the most important in the interaction would allow the design of specific mutations to increase or even disrupt the association in protein–protein complexes. In protein–ligand docking, this knowledge is essential to suggesting chemical modifications on ligand structures guided by the binding site residues. Accordingly, several approaches have been developed.[35,36]

The MM-PB(GB)SA method, as implemented in AmberTools, has all the terms already pairwise decomposed, as a consequence of the double summations in eq 7. On the other hand, in MM-ISMSA the solvation term entails a single summation (eq 14), so the individual solvation for each residue is obtained but not the corresponding contribution of a given interacting pair. The simple addition of the individual solvation values would overestimate this contribution, as not only the atoms involved in the interaction take part in its calculation but also those from their environments. Therefore, we have devised a weighting scheme (see Methods) by means of which the solvation contribution of a given pair is balanced by the sum of the van der Waals and Coulombic interaction in which any of the two residues in that pair is involved.

3. METHODS

3.1. Nonelectrostatics Calculations. We refer here to the calculations involving the van der Waals and the solvent accessible-related terms to account for shape complementarity and the loss in surface area produced upon complex formation (the nonpolar part of the desolvation), respectively. van der Waals interactions (ΔG_{vdW}) are calculated through the well-known 12–6 Lennard-Jones potential:

$$\Delta G_{\text{vdW}} = \sum_{ij} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right]$$

where A_{ij} and B_{ij} are the van der Waals parameters of the atom types to which atoms i and j belong, and r_{ij} is the distance between the ith atom in the protein and the jth atom from the ligand. A_{ij} and B_{ij} parameters were taken directly from the

AMBER ff03 force field.[37,38] The nonpolar part of the desolvation (ΔG_{np}) was modeled as a linear relationship to the change in SASA:

$$\Delta G_{np} = a + b \cdot \Delta \mathrm{SASA}$$

where a is 0.092 kcal·mol^{-1}, b is 0.00542 kcal·mol^{-1} Å^{-2}, and the change in SASA refers to the complex SASA minus the sum of that of the protein and the ligand alone. SASA values were obtained with our own implementation of the LCPO approximation.[39]

Both terms were added to the ISM electrostatic scoring function to configure a more complete tool termed MM-ISMSA, in clear allusion to MM-PB(GB)SA and related approaches.

3.2. Electrostatic Calculations. In all of the cases where PE was employed, we refer to the linearized Poisson−Boltzmann equation; that is, the Boltzmann part of the equation (related to ions in solution) was not taken into account. Atomic radii were automatically assigned with the *tleap* module in AMBER 10 so that they correspond to the "modified Bondi" set.[40]

3.2.1. APBS. APBS uses the adaptive finite element method to solve the Poisson−Boltzmann equation numerically.[30] First, grid size, grid center, and the number of grid points were computed with the *psize.py* module provided in the APBS package, which properly fits the input complexes into their respective grid boxes. Then, the following parameters were chosen: (a) dielectric constants of 4 and 80 for solute and solvent, respectively, (b) a dielectric boundary calculated using a solvent probe radius of 1.4 Å, (c) potentials at the grid points delimiting the box calculated using the multiple Debye−Hückel method, and (d) multigrid PB calculations configured to run in automatic mode. All of these calculations were performed with the APBS program (see section 2.2).

3.2.2. PB/GB. For PB calculations, each complex was immersed in a cubic box with a grid spacing of 0.5 Å. The solute dielectric constant was set to 4, while that of the solvent was set to 80, and the dielectric boundary was calculated using a solvent probe radius of 1.4 Å. The potentials at the grid points delimiting the box were calculated analytically by treating each charge atom as a Debye−Hückel sphere. Similar parameters were employed for GB calculations, namely, the internal and external dielectric constants and the solvent probe radius. These calculations were performed with *mm_pbsa.pl* and *MMPBSA.py* scripts as provided in the AMBER package.

3.2.3. ISM. ISM models the screening effect due to the solvent by means of a sigmoidal distance-dependent dielectric function (eq 12). Solvent-related parameters are the slope of the sigmoidal dielectric function (λ), which has two values, 0.013 for all of the atoms except for those with a formal positive charge and 0.007 for the latter ones; ε, the dielectric constant of the bulk solvent (80); and the solvent probe radius (1.4 Å) to calculate the SASA. For additional parameters, the reader is referred to the original publications.[19,32] All of these calculations were performed with our in-house version of the ISM program (eq 14).

3.3. Hydrogen Bonding Term. The existence of a hydrogen bond was characterized by defining the three atoms involved in the interaction (donor, D; acceptor, A; and the proper hydrogen atom, H) plus the atom bonded to A (X) and three geometrical parameters describing the relative disposition of these atoms: (1) the A···H distance (r), (2) the D···H···A angle (α), and (3) the H···A···X angle (β).

As a training set, we used the Astex Diverse Set (ADS) of protein−ligand complexes. On the other hand, the test sets included the 23 protein−ligand complexes used in the original ISM paper and the 15 protein−protein complexes previously described in section 3.4. Each complex in the training set was visually inspected in PyMOL to determine the number and geometrical parameters of all of the possible hydrogen bonds, whereas for the first test set the program LIGPLOT,[41] as implemented in the PDBsum web server,[42] was used. The first test allowed us to check to what extent the parameters derived for hydrogen bonds in protein−ligand complexes could be extended to protein−protein assemblies. As a default, r was set to 3.5 Å and α to 90°, and no restrictions were imposed on the β angle. Finally, for the second test set, the hydrogen bonds obtained with the HBPLUS program,[43] employing default parameters, were used for comparison.

The distributions of r, α, and β values were analyzed by means of the nonparametrical BOX plot statistical technique. A set of ideal values and upper and lower limits were defined for each variable. These values and the shape of the data distribution were incorporated into *ad hoc* block functions to determine the contribution of each parameter (eq 17) to the total score for each hydrogen bond (HBScore(i), eq 18):

$$\mathrm{score}(x) = \begin{cases} 1 & \text{if } x_{\min\,-\mathrm{ideal}} \le x \le x_{\max\,-\mathrm{ideal}} \\ 1 - \dfrac{x_{\min\,-\mathrm{ideal}} - x}{x_{\min\,-\mathrm{ideal}} - x_{\min}} & \text{if } x_{\min} \le x \le x_{\min\,-\mathrm{ideal}} \\ 1 - \dfrac{x - x_{\max\,-\mathrm{ideal}}}{x_{\max} - x_{\max\,-\mathrm{ideal}}} & \text{if } x_{\max\,-\mathrm{ideal}} \le x \le x_{\max} \\ 0 & \text{if } x_{\max} < x \\ 0 & \text{if } x > x_{\min} \end{cases} \tag{17}$$

where x refers to r, α, and β, and ideal and max values are those obtained from the BOX plot analysis.

$$\mathrm{HBScore}(i) = \prod_{x} \mathrm{score}(x) \tag{18}$$

Hydrogen bonds were further classified by the type of interactions in charged−charged (cc), neutral−charged and charged−neutral (nc), and neutral−neutral (nn) and assigned a numerical value of −3, −2, and −1 kcal/mol, respectively. Finally, HBScore(i) was used to weigh the interaction energy for each hydrogen bond to configure ΔG_{HB}:

$$\Delta G_{HB} = \sum_{i=1}^{NHB} \mathrm{HBScore}(i)\, E_{HB}(i) \tag{19}$$

where i stands for each hydrogen bond; NHB is the total number of hydrogen bonds; and $E_{HB}(i)$ is equal to −3, −2, or −1 depending on the type of hydrogen bond.

3.4. MM-ISMSA Scoring Function. According to the terms described in the above sections (3.1, 3.2.3, and 3.3), the starting equation for MM-ISMSA (eq 20) reads as follows:

$$\Delta G_{\mathrm{binding}} = \Delta G_{\mathrm{vdW}} + \Delta G_{\mathrm{elec}} + \Delta G^{\mathrm{R}}_{\mathrm{desolv}} + \Delta G^{\mathrm{L}}_{\mathrm{desolv}} + \Delta G_{\mathrm{apo}} + \Delta G_{\mathrm{HB}} \tag{20}$$

3.5. Comparison between Methods. A validation test set consisting of 15 antigen−antibody complexes with available 3D structures (PDB ID codes: 1AHW, 1BGX, 1BJ1, 1BVK, 1DQJ, 1FSK, 1I9R, 1IQD, 1JPS, 1KXQ, 1MLC, 1NCA, 1NSN, 1VFB, and 2E6J) was used to compare the performance between MM-ISMSA and the other methods in terms of interaction energies (total, individual terms, and pairwise decomposed). Up to 100 docking poses for each complex were obtained using program

FRODOCK[44] for a total of 1242 structures. Then, the following protocol was employed for each single complex: (a) The AMBER ff03 force field was used to assign atom types and partial charges to each atom in the complexes. (b) Hydrogen atoms were added using the *tleap* module from the AMBER suite assuming standard protonation states for titratable groups. (c) The structures were subjected to an energy refinement process using the GB implicit solvent model as implemented in *sander* (500 cycles of steepest descent followed by 1000 cycles of conjugate gradient until the root-mean-square value of the potential energy gradient was below 0.1 kcal·mol^{-1}·Å^{-1}) to remove possibly existing steric clashes. (d) APBS, MM-ISMSA, and MM-PB(GB)SA calculations were performed on the refined complexes. Finally, we compared the numerical values for the total free energy of binding and its vdW and electrostatic components (Coulombic and desolvation terms) obtained with the three methods.

3.6. The MM-ISMSA Scoring Function in Protein−Protein Docking. A diverse set of 15 protein−protein complexes with experimentally determined 3D structures was taken from the PDB (PDB ID codes: 2FUE, 2JK6, 2LYN, 2O3B, 2ONE, 2Y43, 3AAB, 3AIK, 3DH9, 3F1R, 3G3G, 3MIO, 3PC6, 3PY2, and 3KF3) and used to test the ability of the MM-ISMSA scoring function to select near-native docking poses from a pool of incorrect solutions (decoys). Each complex was separated into two individual structure files containing the receptor and the ligand. The addition of hydrogen atoms and computation of the protonation state of ionizable groups at pH 6.5 were carried out using the H++ server,[45] which relies on AMBER force-field parameters and finite difference solutions to the Poisson−Boltzmann equation. Then, for every pair, the ClusPro server[46] was used to generate different docking poses, and the 10 best-ranked solutions were selected. These structures (plus the native ones) were energy minimized using *sander*. Finally, MM-ISMSA calculations were performed on these refined structures.

The quality of the ranking provided by the MM-ISMSA scoring function was compared to the quality of the docking poses in terms of the set of common contacts found in the docking pose and the native structure, that is, the contact overlap ($C_{overlap}$; eq 21). Two residues were considered to be in contact if any of their respective atoms were closer than a given cutoff distance, 4 Å in our case:

$$C_{overlap} = \frac{\sum_{ij} C_{ij}^{a} C_{ij}^{b}}{\sqrt{(\sum_{ij} C_{ij}^{a})(\sum_{ij} C_{ij}^{b})}} \in [0, 1] \subset \mathbb{R} \tag{21}$$

where i and j stand for receptor and ligand residues, respectively, while a and b represent the native and decoy structures, respectively. Then, C_{ij}^{a} refers to the contacts in the native structure (1 if the contact between i and j exists, 0 otherwise) and C_{ij}^{b}, to those in the decoy.

As it is more common to assess the structural goodness of a scoring function in terms of the root-mean-square deviation (RMSD) found between a scored solution and the corresponding native structure, we have also employed it for comparative purposes although it has been demonstrated that in some cases this measure is devoid of accuracy. We have termed this parameter $RMSD_L$, and its calculation encompasses two steps: first, the alignment of the receptor structure using the McLachlan algorithm[47] and, second, the evaluation of the RMSD for backbone atoms of the whole superimposed structure. These calculations were performed with the ProFit software.[48]

3.7. Pairwise Decomposition. To calculate the pairwise decomposition of the interaction energy, we have devised a weighting scheme by means of which the solvation contribution of a given pair (polar [eq 22] and nonpolar [eq 23]) is balanced by the sum of other interactions (vdW and Coulombic) involving any of the two residues in that pair:

$$\Delta G_{solv}^{i\leftrightarrow j} = w_{i\leftrightarrow j}^{R} \Delta G_{solv}^{i} + w_{i\leftrightarrow j}^{L} \Delta G_{solv}^{j} \tag{22}$$

$$\Delta G_{binding}^{np} = w_{i\leftrightarrow j}^{R} \Delta G_{binding}^{np,i} + w_{i\leftrightarrow j}^{L} \Delta G_{binding}^{np,j} \tag{23}$$

where i and j are the interacting residues belonging to the receptor and ligand, respectively, and

$$w_{i\leftrightarrow j}^{R(L)} = \frac{\Delta G_{binding}^{vdW,i\leftrightarrow j} + \Delta G_{binding}^{elec,coul,i\leftrightarrow j}}{\sum_{k=1}^{N_L(N_R)} (\Delta G_{binding}^{vdW,i\leftrightarrow (j)k} + \Delta G_{binding}^{elec,coul,i\leftrightarrow (j)k})}$$

Based on this scheme, we have classified receptor−ligand interaction pairs into three types: (a) hydrogen bonding; (b) hydrophobic (which includes ΔG_{vdW} and nonpolar contributions from ΔG_{desolv}^{R}, ΔG_{desolv}^{L}, and ΔG_{apo}); and (c) hydrophilic (which includes ΔG_{elec} and the polar contribution from ΔG_{desolv}^{R}, ΔG_{desolv}^{L}, and ΔG_{apo}). Polar and apolar contributions are calculated according to atom types. Namely, N and O atoms are considered polar and the rest apolar. Therefore, any interacting pair is classified as (a) hydrogen bonding, whenever a hydrogen bonding interaction is detected, independently of the other interactions that may occur; (b) hydrophilic, if the relative weight of the hydrophilic term in the total interaction energy is above 60%; (c) hydrophobic, if the relative weight of the hydrophobic term in the total interaction energy is above 60%; and (d) mixed, if the relative weight of the hydrophilic term in the total interaction energy is found to be between 40% and 60%.

3.8. Computational Performance. In this section, we analyze our scoring function (MM-ISMSA) and the MM-PB(GB)SA method in relation to their implementation within the AMBER suite, including the old version (*mm_pbsa.pl*) as well as the new one (*MMPBSA.py*), and their efficiency. First, we analyze the implementation of the algorithms focusing on the programming language employed and how the calculated data are stored and handled. Then, we estimate the efficiency (or the complexity, T) of the algorithms employing the commonly used asymptotic approach (assuming a very large amount of input data) and the big O notation as the parameter to state the order of running time growth. As the elemental unit function, we consider a single energetic calculation which, depending on the method, is represented by the following functions:

$$T_{MM-ISMSA} = (N_R \cdot N_L) + N_R + N_L \tag{24}$$

$$T_{MM-PB(GB)SA} = (N_R + N_L)^2 + N_R^2 + N_L^2 \tag{25}$$

N_R and N_L stand for the number of atoms in the receptor and ligand, respectively.

3.9. MM-ISMSA Graphical User Interface. All of the functionalities available from the MM-ISMSA code have been implemented in a graphical user interface (GUI), written in the Python programming language. This allows its facile use as a plugin to the popular molecular visualization program PyMOL. The GUI front end uses PyMOL software version 1.2 or higher and requires the NumPy (version C 1.3) module and the portable command-line driven graphing utility gnuplot (version 4.6). The GUI has been prepared to be executed on Linux operating systems. The minimum recommended amount of

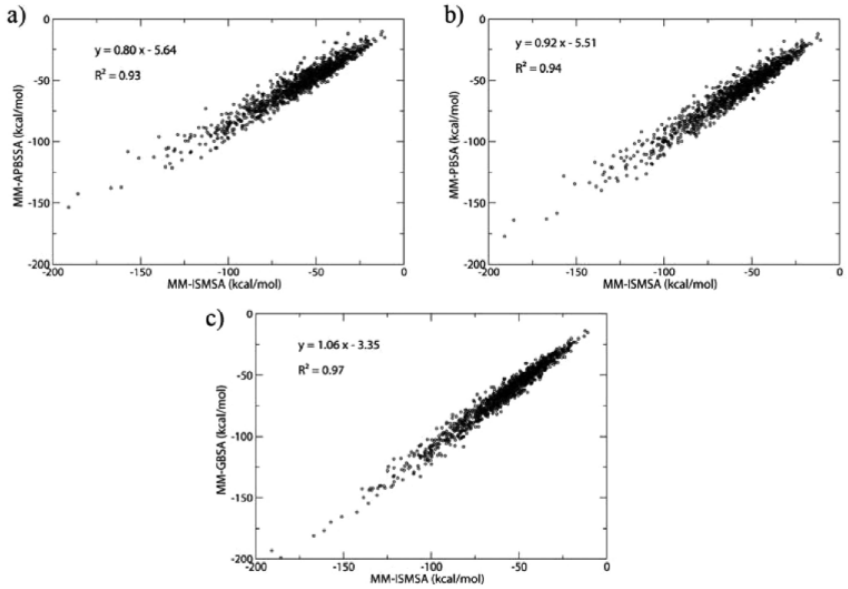


Figure 3. Correlation between total binding free energies ($\Delta G_{binding}$ in kcal/mol) as obtained by (a) MM-APBSSA, (b) MM-PBSA, and (c) MM-GBSA methods and by the MM−ISMSA method.

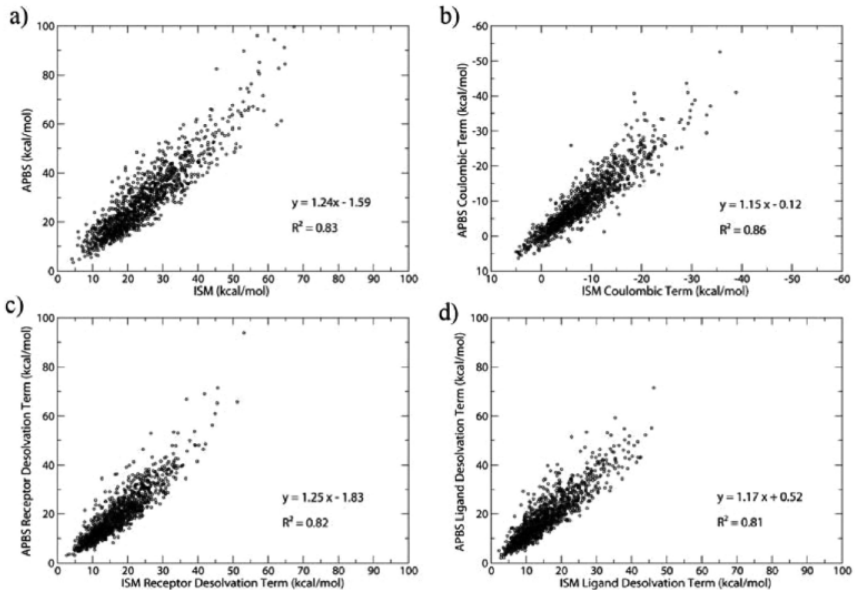


Figure 4. Comparison of total electrostatic binding free energies and their different contributions as obtained by APBS and ISM. (a) Total electrostatic binding free energy. (b) Coulombic contribution. (c) Receptor desolvation. (d) Ligand desolvation.

main memory is around 800 MB (for an average of 900 residues in total per system), and no significant storage capabilities are required to hold the outcomes of the calculations.

4. RESULTS AND DISCUSSION

4.1. MM-ISMSA Compared to MM-APBSSA and MM-PB(GB)SA. In this section, we first compare the numerical values for the total binding free energies ($\Delta G_{binding}$) as obtained from MM-ISMSA, MM-APBSSA, and MM-PB(GB)SA methods. We find MM-ISMSA reproduces MM-APBSSA's $\Delta G_{binding}$ with great accuracy ($r^2 = 0.93$, Figure 3a) as it does when compared to MM-PBSA ($r^2 = 0.94$, Figure 3b) or MM-GBSA ($r^2 = 0.97$, Figure 3c). In addition, we obtain small deviations from MM-PBSA (slope = 0.92) and MM-GBSA (slope = 1.06) and slightly higher from APBS (slope = 0.80). Finally, when MM-PBSA is compared to MM-GBSA in terms of the total interaction energy, the correlation coefficient (r^2 = 0.95), slope (0.86), and intercept (−3.16) are similar to the values obtained when comparing MM-

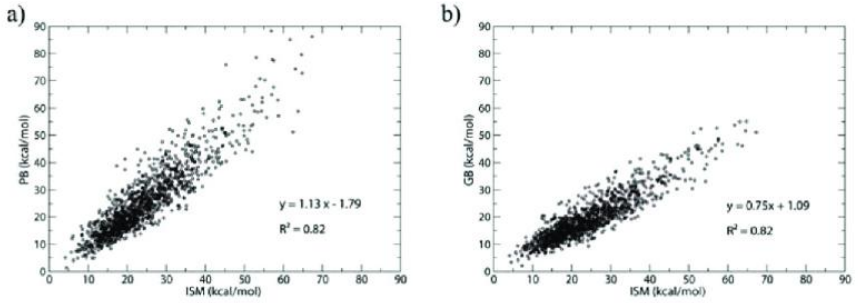


Figure 5. Comparison of total electrostatic binding free energies obtained by ISM with those calculated using (a) PB (as implemented in MM-PBSA) and (b) GB (as implemented in MM-PBSA).

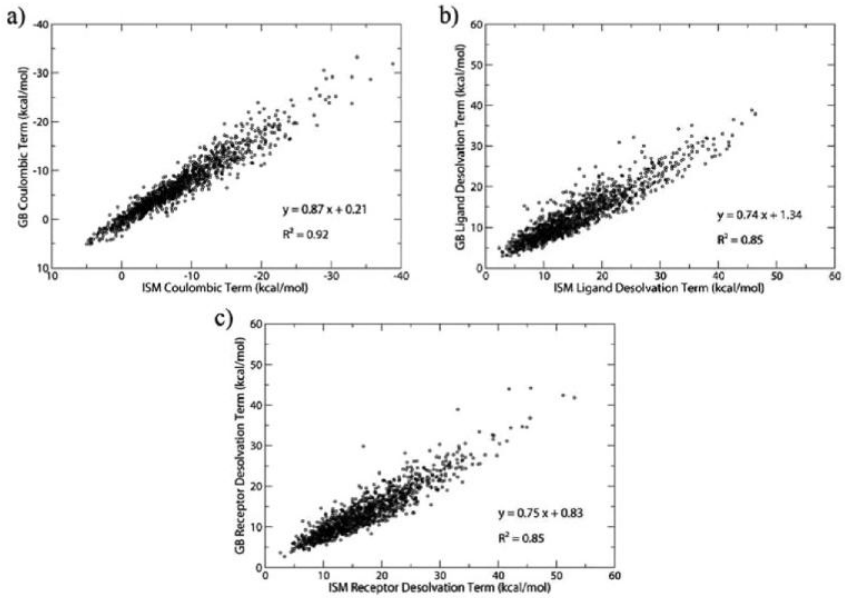


Figure 6. Comparison between GB and ISM in terms of the different contributions to the total electrostatic binding free energy as obtained by applying the three-calculation method. (a) Coulombic contribution. (b) Ligand desolvation. (c) Receptor desolvation.

ISMSA to either MM-GBSA or MM-PBSA. As the non-electrostatic terms (ΔG_{vdW} and ΔG_{np}) are always computed in the same way, the rest of this section will focus on the electrostatic part ($\Delta G^{elec}_{binding}$) of $\Delta G_{binding}$ and its decompositions into the Coulombic ($\Delta G^{elec,coul}_{binding}$) and receptor and ligand desolvation terms ($\Delta G^{elec,desolv_R}_{binding}$ and $\Delta G^{elec,desolv_L}_{binding}$, respectively). In the following, unless otherwise stated, we will refer to the methods simply as APBS, ISM, GB, or PB.

The correlation between APBS and ISM in terms of $\Delta G^{elec}_{binding}$ is very good (r^2 = 0.83, Figure 4a), and the same is true when individual components are considered: $\Delta G^{elec,coul}_{binding}$ (r^2 = 0.86, Figure 4b), $\Delta G^{elec,desolv_R}_{binding}$ (r^2 = 0.82, Figure 4c), and $\Delta G^{elec,desolv_L}_{binding}$ (r^2 = 0.81, Figure 4d). Slopes are close to unity (1.15 for $\Delta G^{elec,coul}_{binding}$ and 1.17 for $\Delta G^{elec,desolv_L}_{binding}$) or slightly higher (1.24 and 1.25 for $\Delta G^{elec}_{binding}$ and $\Delta G^{elec,desolv_R}_{binding}$, respectively), and the intercepts are very small in all cases.

Furthermore, when ISM is compared to either PB (Figure 5a) or GB (Figure 5b), the correlation coefficients (r^2 = 0.82 in both cases), slopes (1.13 and 0.75), and intercepts (1.79 and 1.09) are on the same order as before.

As stated above, the thermodynamic cycle on which ISM is based (Figure 2) directly dissects, by construction, the electrostatic contribution to binding into its components (eq 13) in just a single step. On the contrary, three calculations are needed to obtain the same partition scheme when PB or GB models (eqs 9–11) are employed. In particular, when the triple calculation is used, ISM compared to GB (Figure 6) affords a correlation coefficient close to 0.9 (r^2 = 0.85) for both desolvations (Figure 6b,c), while the slopes indicate that these values are overestimated by the GB model (0.75 and 0.74 for receptor and ligand desolvation, respectively). The agreement between Coulombic terms is even better, with both the correlation coefficient and the slope yielding a value of ~0.9 (Figure 6a).

4.2. The MM-ISMSA Scoring Function in Protein–Protein Docking. We have challenged the MM-ISMSA scoring function by mixing the native structures for a set of 15 protein–protein complexes with docking decoys (10 for each complex obtained from the ClusPro program) and then evaluating the binding free energy for each complex (eq 14). Although this

function provides more detailed information on the binding event than just the global value of the binding energy (hydrogen bond types and count, HBScore, van der Waals and electrostatic interactions, desolvation terms, etc.), we were not able, however, to discriminate between the native structure and the decoys on the basis of this detailed individual information. Rather, it was only the global binding energy parameter which really selects the native pose as the best scored pose among the decoys in 12 out of the 15 complexes studied here (80% of success, Figure 7).

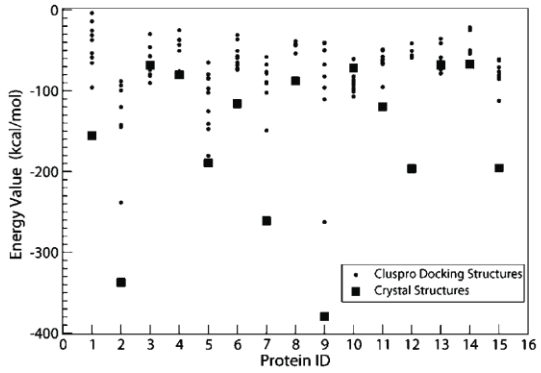


Figure 7. Total free energies of binding for each of the decoys and the X-ray structure in the test set. The small black dots represent the decoys, while the black squares symbolize the X-ray structure.

The finding that descriptors as important as the types and number of hydrogen bonds or the electrostatic complementarity between receptor and ligand, considered as essential in protein−protein recognition, are not able to discriminate a native structure from a pool of decoys might be due to the ability of ClusPro to provide challenging docking poses, although with a significant difference in contact patterns (Figure 8a). In fact, less than half of the complexes showed $C_{overlap}$ values above 0.7. On the contrary, ClusPro is able to supply in most of the cases (10 out of 15) a decoy closer than 10 Å to the native structure (Figure 8b).

Finally, we have a weak correlation ($r^2 \sim 0.5-0.6$) between the rankings provided by MM-ISMSA and the contact overlap (Figure 9a) or $RMSD_L$ (Figure 9b). Despite these modest figures it is worth noting that in many cases (10 out of 12 if complexes without very low $C_{overlap}$ values are excluded) good docking poses according to the MM-ISMSA scoring function usually correspond to an assembly with high and low $C_{overlap}$ and $RMSD_L$ values, respectively.

4.3. Pairwise Decomposition. Figure 10 shows the relationship between all individual inter-residue interaction energies calculated by MM-ISMSA and MM-GBSA for the 15 proteins in the second test set. A very good correlation was obtained ($r^2 = 0.95$), indicating the feasibility of employing MM-ISMSA as an alternative to MM-GBSA.

4.4. Computational Performance. In routine calculations, when dealing with relatively large protein−protein complexes, we have found a tremendous bottleneck in the use of the mm_pbsa.pl module in AMBER to obtain the partition of the interaction energy into pairwise residue contributions. In fact, this decomposition scheme is actually impracticable in many cases due to the large execution times required, not to mention the huge demands on memory and disk space. After some careful examination, we concluded that the main culprit for this appeared to be the use of *hash tables* to store the data. On the other hand, we were aware that *perl*, being an interpreted language, is not adequate to handle highly computationally demanding mathematical calculations. To circumvent these problems, we decided to rewrite in C programming language the main part of the *mm_pbsa.pl* module where the statistical calculations are performed (*mm_pbsa_statistics.pm* module), replacing the *hash* tables with single *arrays*. We refer to this new code as optimized *mm_pbsa.pl*. By the time we were developing this optimized code, a new version of the *mm_pbsa.pl* module was released under the name *MMPBSA.py*. This new version avoids some problems of the early module by executing sander binaries on each snapshot and using a more adequate data structure and methods from *Python* programming language. Then, we compared the performance of the MM-ISMSA code to the old (*mm_pbsa.pl*) and new (*MMPBSA.py*) modules implemented in the AMBER package and to the C optimized module (Figure 11) in terms of execution times.

Three main observations can be derived from Figure 11: First, in the limit of high number of residues, the ISM code achieves execution times 4 orders of magnitude smaller than the old *mm_pbsa.pl* module. Second, at the same limit, the new implemented *MMPBSA.py* and our *mm_pbsa.pl* optimized module show very close execution times. Third, ISM outperforms all the other codes tested here, irrespective of the number of residues used in the calculation.

Continuing with the pairwise residue decomposition, and according to eqs 24 and 25 above, it is relatively straightforward to realize that the efficiency of both algorithms grows on the

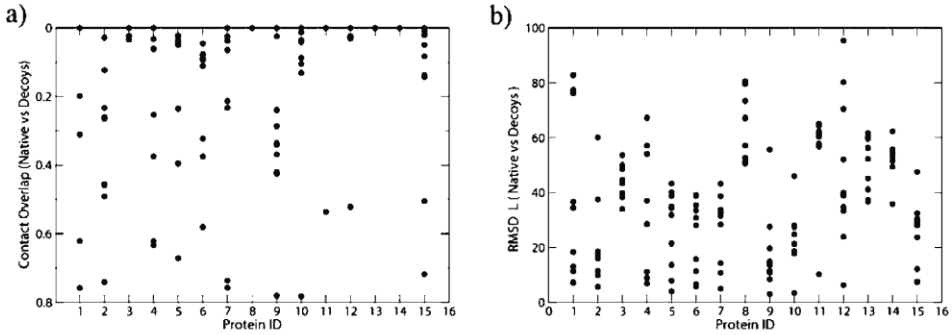


Figure 8. (a) Contact overlap (eq 21) and (b) $RMSD_L$ for the decoys generated with ClusPro for each of the targets in the test set.

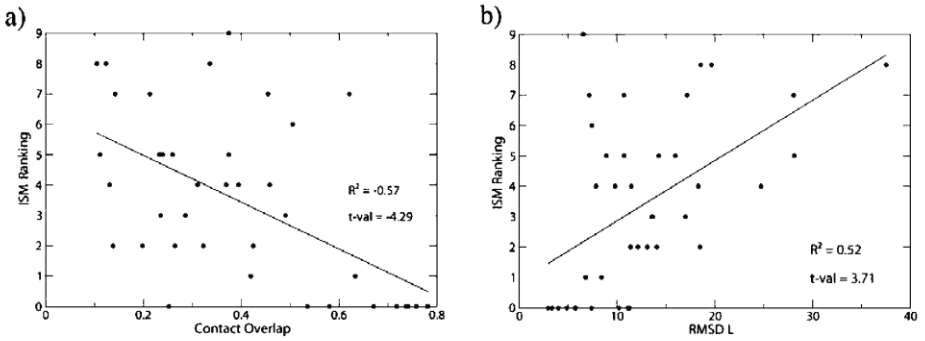


Figure 9. Relationships between (a) contact overlap (eq 21) and (b) $RMSD_L$ and ISM ranking based on the total binding energy for the decoys generated with ClusPro for each of the targets in the test set.

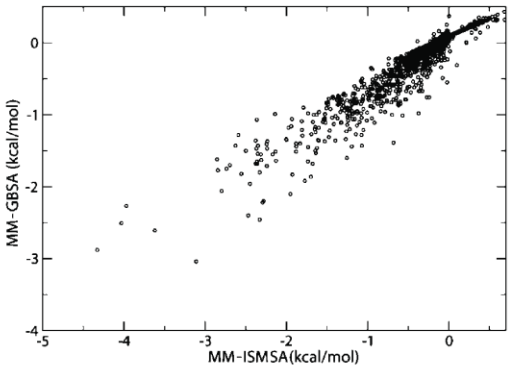


Figure 10. Relationship between the pairwise-decomposed binding energies obtained by MM-ISMSA and MM-GBSA methods for the 15 proteins in the second test set.

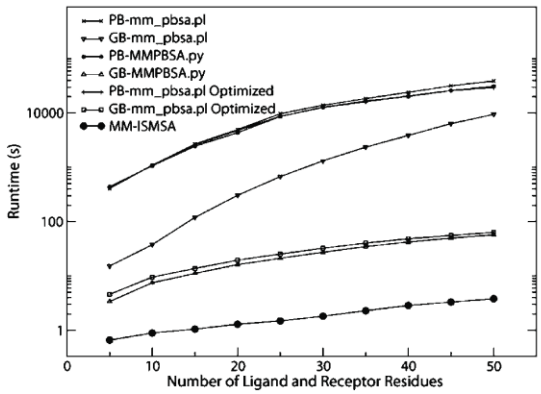


Figure 11. Run time (log scale) dependence on the number of residues for the different methods compared in this study.

order of $O(N^2)$, provided that N_R and N_L are of the same magnitude. However, the number of operations to be done in MM-ISMSA is smaller than in MM-PB(GB)SA, and this is due to the double counting of the cross-interaction terms between the receptor and the ligand in the latter method (double summation in eq 7, ΔG_{pol} term, with the indexes running from 1 to N for both receptor and ligand, respectively), while in the former they are calculated just once (single summation in the second term on the right-hand side of eq 14 with a single index running from 1 to the total number of atoms in the complex). To analyze in more detail the relative performance of both algorithms, we studied the two most representative cases: (a) protein−protein docking assuming an equal number of residues for both the receptor and ligand and (b) protein−ligand docking, where the number of ligand residues (one) can be neglected as compared to the number of residues in the receptor. Then, in the case of protein−protein docking:

$$\text{if } N_R \approx N_L$$
$$= N, \lim_{N\to\infty} \frac{T_{MM-PB(GB)SA}(N)}{T_{MM-ISMSA}(N)}$$
$$= \lim_{N\to\infty} \frac{6N^2}{N^2 + 2N} = 6 \quad (26)$$

and in the case of protein−ligand docking, we obtain:

$$\text{if } N_R \gg N_L, \lim_{N_R\to\infty} \frac{T_{MM-PB(GB)SA}(N_R)}{T_{MM-ISMSA}(N_R)}$$
$$= \lim_{N_R\to\infty} \frac{N_R^2}{N_R} = \infty \quad (27)$$

Therefore, even though both algorithms scale within the same complexity order for protein−protein interactions (eq 26), the differences are significant mainly due to the fact that MM-ISMSA performs six times less operations than MM-PB(GB)SA. To illustrate this point, consider a protein−protein complex consisting of a dimer with an average size of 150 residues per monomer. Assuming that MM-ISMSA requires the same amount of time to perform a unit calculation as MM-GBSA or MM-PBSA does (around 0.005 and 2 s, respectively), these differences would be translated into 83% or 98% savings in execution time, for protein−protein or protein−ligand complexes, compared to MM-GBSA or MM-PBSA, respectively.

Finally, it is worth commenting that under the assumption that *sander* energies are calculated following an optimized compiled code, and that the statistical optimized code accounts for less than 0.01% of the total execution time, we can conclude that *sander* has reached its optimization limit, and no further improvement can be performed unless a new implementation of the part of the code in charge of the energy calculations is undertaken.

4.5. Hydrogen Bonding. Table 1 contains the geometrical parameters estimated from the statistical analysis of the hydrogen bonds.

Table 1. Geometrical Parameters Defining the Hydrogen Bonding Interactions[a]

	min	min−ideal	max−ideal	max
r	1.5[a]	1.8[a]	2.4	2.7
α	100[b]	130	165[b]	180
β	90	115	145	180

[a] r, in Å, is the distance between the hydrogen and acceptor atoms, and α and β, in degrees, are the angles between donor, hydrogen, and acceptor atoms, and the hydrogen, acceptor, and the atom bound to the acceptor atom, respectively. [b]The final values for these variables were finely tuned to avoid penalizing those hydrogen bonds that, although geometrically plausible, are either statistically underrepresented (α_{min} = 90° and $\alpha_{max\text{-}ideal}$ = 180°) or already penalized by the van der Waals term (r_{min} = 0 and $r_{min\text{-}ideal}$ = 0).

Using these values, we have been able to correctly recover 92.5% of the hydrogen bonds in the training set and 83.7% in the first test set. In most of the cases, the reason why those hydrogen bonds were not identified was either because their defining values were close to the limiting ones or because their geometrical arrangement cast some doubts on their formation. But more importantly, their associated interaction energies were well below the average energies obtained for the recovered hydrogen bonds. Finally, for the second test set, all the hydrogen bonds were identified when compared to HBPLUS as the reference method.

4.6. The PyMOL Plugin. Due to the possible steep learning curve of the application, which requires significant knowledge of the underlying operations, we have implemented a GUI accessible within the popular molecular editor PyMOL. This plugin was designed to execute the application and process the results of the calculations. Figure 12 shows some of the more important graphic capabilities of the plugin, while a complete description is included in the user's guide available from the Web site.

Once the plugin is invoked from the PyMOL plugin interface, it displays a window with four main sections organized in different tabs: Configure, Run, Global Analysis, and Residue Analysis. At the bottom of this window, an OK button closes the plugin. The Configure tab, which is activated by default, comprises the following fields: (a) an area that contains two user-configurable variables (the path to the ISM executable file and the working directory) and a Save button that allows the user to save this configuration; (b) a second area that depicts the logos of the institutions involved in its development; and (c) a scrollable window with information about the plugin, its authors and institutions involved, contact details, the license, the disclaimer, and plugin update information.

The Run tab has two sections: (a) on the left, the user can (i) select to work with a single structure (either a unique PDB file or two top and crd AMBER-type files) or with an MD trajectory file, (ii) set the type of input (single structure or trajectory file), and (iii) set up the per residue analysis; (b) on the right, the user can (i) set up the root name for the output files, (ii) in the case of an MD trajectory file, select the initial, final, and step size of the snapshot to be analyzed, and (iii) explore the output of the calculation in a small window. A Run button on the right on this window starts the calculation. A Stop button is also provided to stop the calculation.

In the Global Analysis tab, the user can do the following: (a) Load the results from a previous calculation by typing the root of the file name or selecting the file. (b) Export the results of the current calculation. The options are (i) a PDB file with the desolvation energy values per residue loaded into the temperature factor field for an easy visualization in PyMOL, (ii) a file

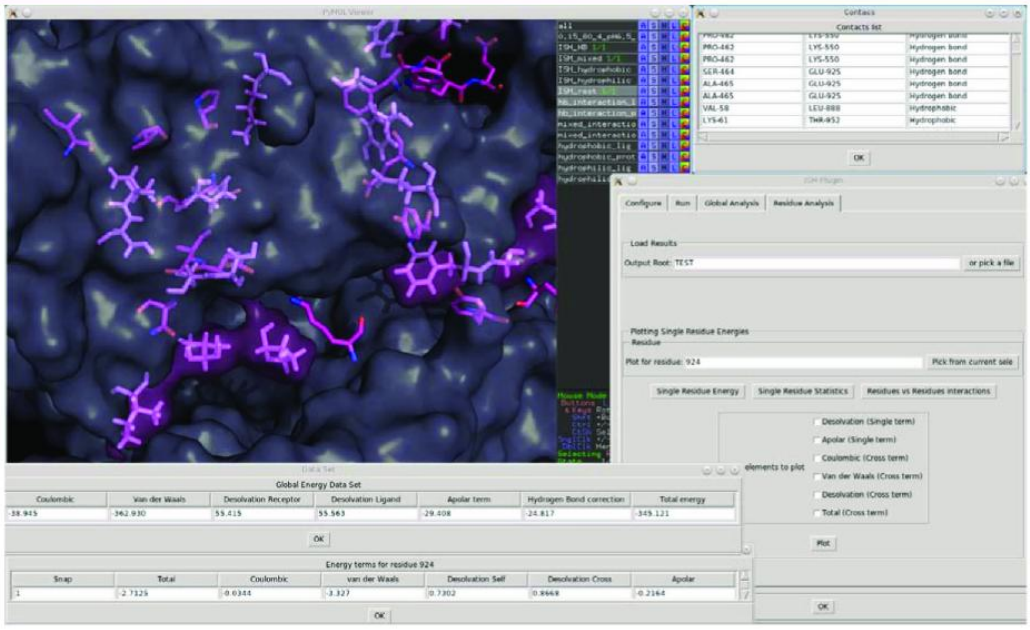


Figure 12. Combined snapshots of some graphical capabilities provided by the PyMOL-implemented MM-ISMSA plugin.

with the Global Energy and its components, and (iii) a file with the Global Statistics in case an MD trajectory is processed. (c) Visualize the contacts between ligand and receptor in PyMOL at various levels: (1) all the contacts, (2) only hydrophobics, (3) only hydrophilics, (4) mixed contacts, and (5) hydrogen bonds. It is worth remembering that the contact classification is not based on the type of side chains found in the interacting residue but on the kind of interaction energy (see section 2.5). (d) Plot the evolution of the different energetic terms along the MD trajectory. These are Coulombic, van der Waals, receptor and ligand desolvation energies, apolar, hydrogen bond, and total.

The last tab, Residue Analysis, allows the user to perform single residue analysis and visualization. The main functionalities available are as follows: (a) Load the results of a previous calculation by typing the root of the file name or selecting the file. (b) Export the results of the current calculation. The options are (i) a file with the energies by residue with its components, (ii) a file with the Global Statistics in case an MD trajectory is processed, and (iii) a matrix with residue−residue interaction energies. (c) Plot the evolution of the different energy terms along the MD trajectory. These are Coulombic (cross term), van der Waals (cross term), receptor and ligand desolvation energies (single and cross terms), apolar (single term), and total (cross term).

Within PyMOL, there are currently some other plugins developed to run and analyze MD trajectories produced with the AMBER suite of programs,[49] to compute molecular electrostatic potentials that can be used as the basis for the estimation of binding and desolvation energies,[30] or to calculate protein−protein interactions.[50] However, they are not as complete as our MM-ISMSA plugin. On the other hand, a great variety of tools are included within the molecular visualization program VMD[51] to analyze MD trajectories calculated with the CHARMm force field. But again, they appear as separate plugins although they are completely integrated within the VMD working environment. In addition, there are some Web-based applications that can estimate contact and binding free energies in protein−protein complexes,[52] predict hot spot residues in protein interfaces,[53] and analyze and visualize contacts at the interface of biomolecular complexes,[42,54] just to mention a few. Finally, and as far as MD protocols are concerned, GUIs for the most commonly used MD codes have appeared recently.[55] Summarizing, we think that our MM-ISMSA PyMOL plugin condenses some of the advantages of the aforementioned tools while maintaining its integrity in a single, unified tool that is implemented in a widely used and powerful molecular graphics program.

■ CONCLUSIONS

A new scoring function for protein−protein docking, MM-ISMSA, which incorporates desolvation and hydrogen bonding terms explicitly, is presented. This function allows calculation in a given protein−protein complex of (i) the total binding free energy, (ii) the contributions from different components, (iii) individual residue desolvations, and (iv) all pairwise residue interactions.

The accuracy of MM-ISMSA was tested using two different protein−protein sets; in the first one, a total of 1242 structures (15 experimentally determined antigen−antibody complexes and 1227 decoys, with a maximum of 100 decoys per complex) were used to study whether or not MM-ISMSA was able to reproduce the interaction energies (i−iv above) as compared to other well-established methods in the field. The results showed that in all cases a good agreement was achieved. The second set (15 diverse experimentally determined complexes and 135 decoys, with a maximum of 10 per complex) was used to test the ability of the MM-ISMSA scoring function to select near-native docking poses from a pool of solutions (decoys). The outcome was an 80% success rate.

Besides its accuracy, an additional advantage of MM-ISMSA is its reduced computational cost, as it is able to analyze large systems (~1000 residues) in less than 5 s, yielding a complete report on the different energy terms and its decomposition. Compared to the commonly used MM-PB(GB)SA method (as implemented in AmberTools), MM-ISMSA performs 6 times fewer calculations than MM-PB(GB)SA. For this reason, it should be particularly preferable to process long MD trajectories.

Finally, MM-ISMSA has been implemented as a plugin for the popular molecular visualization program PyMOL, although it can also be used in command-line mode. The code is open source and is offered free of charge to noncommercial parties for download following registration at the CBM Bioinformatics Unit's web page (http://ub.cbm.uam.es/).

■ AUTHOR INFORMATION

Corresponding Author

*Phone: + 34 911 964 633. Fax: + 34 911 964 422. E-mail: amorreale@cbm.uam.es.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by grants from CICYT (SAF2009-13914-C02-02 to F.G.) and Comunidad Autónoma de Madrid (S-BIO-0214-2006 [BIPEDD] and S2010-BMD-2457 [BIPEDD2] to A.M. and F.G.). A.M. acknowledges financial support from Fundación Severo Ochoa through the AMAROUTO program. J.K., H.G.D.S., and A.N.-S. were supported by grants from Ministerio de Economía y Competitividad (BFU2011-24595) and Ministerio de Educación y Ciencia (CSD2006-00023). R.G.-R. enjoyed a MICINN contract from "Programa de Personal Técnico y de Apoyo 2008", and Á.C.-C. is the recipient of FPU grant AP2009-0203 from Ministerio de Educación. We thank Dr. Pablo Chacón (IQFR-CSIC, Madrid) for providing us with the structures used in the validation test set and Dr. Ugo Bastolla for useful discussions and encouragement.

■ REFERENCES

(1) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813−1818.

(2) Kortemme, T.; Joachimiak, L. A.; Bullock, A. N.; Schuler, A. D.; Stoddard, B. L.; Baker, D. Computational redesign of protein-protein interaction specificity. *Nat. Struct. Mol. Biol.* **2004**, *11*, 371−379.

(3) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **2002**, *47*, 409−443.

(4) Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J. Med. Chem.* **2010**, *53*, 8461−8467.

(5) Ben-Shimon, A.; Eisenstein, M. Computational mapping of anchoring spots on protein surfaces. *J. Mol. Biol.* **2010**, *402*, 259−277.

(6) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of folded proteins. *Nature* **1977**, *267*, 585−590.

(7) Jorgensen, W. L. Efficient drug lead discovery and optimization. *Acc. Chem. Res.* **2009**, *42*, 724−733.

(8) Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.;

Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossvry, I.; Klepeis, J. L.; Layman, T.; McLeavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. Anton, a special-purpose machine for molecular dynamics simulation. *SIGARCH Comput. Archit. News* **2007**, *35*, 1−12.
(9) Top 500 Supercomputer sites. http://www.top500.org/ (accessed July 20, 2012).
(10) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334*, 517−520.
(11) Kruse, A. C.; Hu, J.; Pan, A. C.; Arlow, D. H.; Rosenbaum, D. M.; Rosemond, E.; Green, H. F.; Liu, T.; Chae, P. S.; Dror, R. O.; Shaw, D. E.; Weis, W. I.; Wess, J.; Kobilka, B. K. Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* **2012**, *482*, 552−556.
(12) Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. How does a drug molecule find its target binding site? *J. Am. Chem. Soc.* **2011**, *133*, 9181−9183.
(13) Orozco, M.; Luque, F. J. Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems. *Chem. Rev.* **2000**, *100*, 4187−4226.
(14) Roux, B.; Simonson, T. Implicit solvent models. *Biophys. Chem.* **1999**, *78*, 1−20.
(15) Honig, B.; Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **1995**, *268*, 1144−1149.
(16) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.
(17) Kuhlman, B.; Baker, D. Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 10383−10388.
(18) Yin, S.; Biedermannova, L.; Vondrasek, J.; Dokholyan, N. V. MedusaScore: an accurate force field-based scoring function for virtual drug screening. *J. Chem. Inf. Model.* **2008**, *48*, 1656−1662.
(19) Morreale, A.; Gil-Redondo, R.; Ortiz, A. R. A new implicit solvent model for protein-ligand docking. *Proteins* **2007**, *67*, 606−616.
(20) Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668−1688.
(21) *The PyMOL Molecular Graphics System*, version 1.1r2pre; Schrödinger, LLC: Portland, OR, 2008.
(22) Gilson, M. K.; Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21−42.
(23) Aqvist, J.; Medina, C.; Samuelsson, J. E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385−391.
(24) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., 3rd. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889−897.
(25) Gilson, M. K.; Honig, B. Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins* **1988**, *4*, 7−18.
(26) Zhou, H. X.; Gilson, M. K. Theory of free energy and entropy in noncovalent binding. *Chem. Rev.* **2009**, *109*, 4092−4107.
(27) Warwicker, J.; Watson, H. C. Calculation of the electric potential in the active site cleft due to α-helix dipoles. *J. Mol. Biol.* **1982**, *157*, 671−679.
(28) Baker, N.; Holst, M.; Wang, F. Adaptive multilevel finite element solution of the Poisson−Boltzmann equation II. Refinement at solvent-accessible surfaces in biomolecular systems. *J. Comput. Chem.* **2000**, *21*, 1343−1352.
(29) Rashin, A. A.; Namboodiri, K. A simple method for the calculation of hydration enthalpies of polar molecules with arbitrary shapes. *J. Phys. Chem.* **1987**, *91*, 6003−6012.
(30) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037−10041.
(31) Mehler, E. L., The Lorentz-Debye-Sack Theory and Dielectric Screening of Electrostatic Effects in Proteins and Nucleic Acids. In *Molecular Electrostatic Potential: Concepts and Applications*; Murray, J. S.; Sen, K., Eds.; Elsevier Science: Amsterdam, 1996; Vol. 3, pp 371−405.
(32) Hassan, S. A.; Guarnieri, F.; Mehler, E. L. A General Treatment of Solvent Effects Based on Screened Coulomb Potentials. *J. Phys. Chem. B* **2000**, *104*, 6478−6489.
(33) Hassan, S. A.; Mehler, E. L.; Zhang, D.; Weinstein, H. Molecular dynamics simulations of peptides and proteins with a continuum electrostatic model based on screened Coulomb potentials. *Proteins* **2003**, *51*, 109−125.
(34) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.
(35) Luque, F. J.; Curutchet, C.; Munoz-Muriedas, J.; Bidon-Chanal, A.; Soteras, I.; Morreale, A.; Gelpi, J. L.; Orozco, M. Continuum solvation models: Dissecting the free energy of solvation. *Phys. Chem. Chem. Phys.* **2003**, *5*, 3827−3836.
(36) Morreale, A.; de la Cruz, X.; Meyer, T.; Gelpi, J. L.; Luque, F. J.; Orozco, M. Partition of protein solvation into group contributions from molecular dynamics simulations. *Proteins* **2005**, *58*, 101−109.
(37) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999−2012.
(38) Lee, M. C.; Duan, Y. Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model. *Proteins* **2004**, *55*, 620−634.
(39) Weiser, J.; Shenkin, P. S.; Still, W. C. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *J. Comput. Chem.* **1999**, *20*, 217−230.
(40) Tsui, V.; Case, D. A. Theory and applications of the generalized born solvation model in macromolecular simulations. *Biopolymers* **2000**, *56*, 275−291.
(41) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.* **1995**, *8*, 127−134.
(42) Laskowski, R. A. PDBsum new things. *Nucleic Acids Res.* **2009**, *37*, D355−359.
(43) McDonald, I. K.; Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **1994**, *238*, 777−793.
(44) Garzon, J. I.; Lopez-Blanco, J. R.; Pons, C.; Kovacs, J.; Abagyan, R.; Fernandez-Recio, J.; Chacon, P. FRODOCK: a new approach for fast rotational protein-protein docking. *Bioinformatics* **2009**, *25*, 2544−2551.
(45) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res.* **2005**, *33*, W368−371.
(46) Comeau, S. R.; Gatchell, D. W.; Vajda, S.; Camacho, C. J. ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Res.* **2004**, *32*, W96−99.
(47) McLachlan, A. Rapid comparison of protein structures. *Acta Crystallogr., Sect. A* **1982**, *38*, 871−873.
(48) Martin, A. C. R.; Porter, C. T. *ProFit*, version 3.1; University College London: London, 2009.
(49) Lill, M.; Danielson, M. Computer-aided drug design platform using PyMOL. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 13−19.
(50) Steinkellner, G.; Rader, R.; Thallinger, G. G.; Kratky, C.; Gruber, K. VASCo: computation and visualization of annotated protein surface contacts. *BMC Bioinf.* **2009**, *10*, 32.
(51) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **1996**, *14* (33−38), 27−38.
(52) Camacho, C. J.; Zhang, C. FastContact: rapid estimate of contact and binding free energies. *Bioinformatics* **2005**, *21*, 2534−2536.
(53) Assi, S. A.; Tanaka, T.; Rabbitts, T. H.; Fernandez-Fuentes, N. PCRPi: Presaging Critical Residues in Protein interfaces, a new

computational tool to chart hot spots in protein interfaces. *Nucleic Acids Res.* **2010**, *38*, e86.
(54) Vangone, A.; Spinelli, R.; Scarano, V.; Cavallo, L.; Oliva, R. COCOMAPS: a web application to analyze and visualize contacts at the interface of biomolecular complexes. *Bioinformatics* **2011**, *27*, 2915–2916.
(55) Knapp, B.; Schreiner, W. Graphical user interfaces for molecular dynamics-quo vadis? *Bioinform. Biol. Insights* **2009**, *3*, 103–107.

4.5.3.- Implementación de un optimizador de grados de libertad torsionales en el programa de *docking* CRDOCK

La contribución de la autora de esta tesis al artículo (Cortés Cabrera et al., 2012) fue, además de la incorporación del término de HB en la función de energía de MM-ISMSA explicado en el apartado anterior, el llevar a cabo la implementación de un algoritmo para minimizar la energía de los ligandos dentro del centro activo de las proteínas durante el refinado de las soluciones de *docking*, usando sus ángulos de torsión como grados de libertad y partiendo nuevamente del conjunto diverso de Astex para la evaluación de los resultados. Durante la generación de confórmeros previa al *docking,* el programa ALFA (Gil Redondo R, 2006) identifica automáticamente los enlaces rotables y les asigna su estado rotamérico en base a la hibridación de los átomos involucrados. La asignación de los ángulos torsionales o diedros (determinados por 4 átomos) viene definida a partir de un conjunto de reglas químicas específicas según los grupos funcionales existentes a ambos lados del enlace rotable. Esta aproximación tiene sus limitaciones ya que los complejos proteína-ligando pueden estar muy empaquetados, presentando desviaciones con respecto a los ángulos ideales de un ligando aislado. Esto implica que el programa ALFA no será capaz de reconstruir la conformación cristalográfica durante la generación de confórmeros previa al *docking,* por lo que no está garantizado que, dentro de las conformaciones propuestas por el algoritmo, aparezca alguna cercana a la que el ligando adopta en el sitio activo.

En la fase de evaluación hemos observado que si se usa la conformación del ligando unido a la proteína que se encuentra en el PDB (*docking* rígido), nuestro programa de *docking* coloca correctamente (*i.e.* el RMSD entre la pose cristalográfica y la solución de *docking* es menor de 2.0 Å) el 93% de los ligandos del conjunto de Astex. Esto nos indica que la función de *scoring* del programa de *docking* funciona bien siempre y cuando, durante el muestreo conformacional del ligando, encontremos aquella cercana a la nativa. Sin embargo al realizar el *docking* flexible a partir de los confórmeros del ligando pre-generados con ALFA y usando la versión anterior de nuestro algoritmo de refinado donde solo permitíamos movimientos de cuerpo rígido, la fiabilidad de los resultados se reducía significativamente (entorno al 55%) debido a la ausencia de una conformación adecuada entre las propuestas. La minimización de la energía libre del complejo proteína-ligando mediante la aplicación de torsiones en los enlaces rotables del ligando en el contexto del sitio activo nos permite mejorar la complementariedad ligando-proteína y su energía de interacción en el 88% de los complejos mientras que el 12% restante ésta se mantiene constante. Conseguimos un RMSD más bajo en el 20% de los complejos (acumulando mejoras de más de 1 Å de RMSD respecto al *docking* sin

el minimizador torsional) pero también un RMSD más alto en el 10 % de los complejos. En resumen, tras la minimización de los ángulos torsionales del ligando el 61% de los complejos presentaron una solución buena de *docking* (*i.e.* RMSD < 2.0 Å) con respecto al ligando cristalográfico (ejemplo en la Figura 26). A pesar de las mejoras observadas a nivel de RMSD, no en todos los casos fue suficiente para ser consideradas buenas en *docking*. Por otra parte, no observamos correlación entre la energía de unión estimada y el RMSD de la pose asociada.

Hemos demostrado que la inclusión de los DOFs torsionales del ligando durante la minimización de la energía de las poses de *docking* a través de la función de *scoring* MM-ISMSA mejora la precisión de los resultados respecto a la versión anterior del protocolo que usaba un conjunto de confórmeros rígidos pre-calculados con el programa ALFA. Con la inclusión de la flexibilidad de los enlaces rotables del ligando, un solo confórmero es suficiente para alcanzar un muestreo adecuado y soluciones de *docking* más cercanas a las cristalográficas manteniendo un tiempo de cálculo por *docking* reducido.

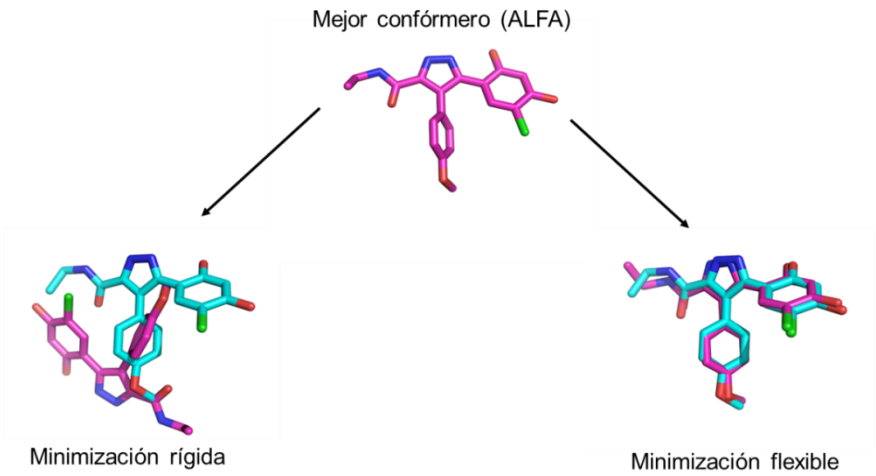


Figura 26. Ejemplo de la influencia de la minimización energética de la soluciones de *docking* (en rosa) respecto a la pose del ligando cristalográfico (en azul) a partir del confórmero generado con el programa ALFA con la menor energía. [PDB ID: 2BSM (Dymock et al., 2005)]. El RMSD tras la minimización rígida (rotaciones y traslaciones) alcanza 6.7 Å, mientras que el RMSD tras la minimización incluyendo los DOFs torsionales se reduce hasta 0.6 Å. Los átomos de hidrógeno no se muestran en la imagen por claridad.

Los datos presentados en esta introducción respecto a la mejora en la capacidad predictiva del programa de *docking* (61%) hacen referencia a la influencia de la nueva función de energía MM-ISMSA con el término de HB en combinación con la minimización de los DOFs torsionales del ligando usando el programa previo (CDOCK). Sin embargo, con la incorporación de éstas y otras mejoras fuera del contexto de esta tesis, la capacidad predictiva del nuevo programa presentado (CRDOCK) asciende al 73% para el conjunto diverso de Astex (RMSD < 2 Å).

Artículo 6

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

CRDOCK: An Ultrafast Multipurpose Protein−Ligand Docking Tool

Álvaro Cortés Cabrera,[†,‡] Javier Klett,[‡] Helena G. Dos Santos,[‡] Almudena Perona,[‡,§] Rubén Gil-Redondo,[‡,§] Sandrea M. Francis,[∥] Eva M. Priego,[⊥] Federico Gago,[†] and Antonio Morreale*[,‡]

[†]Departamento de Farmacología, Universidad de Alcalá, E-28871 Alcalá de Henares, Madrid, Spain

[‡]Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa (CSIC-UAM), Campus UAM, c/Nicolás Cabrera 1, E-28049 Madrid, Spain

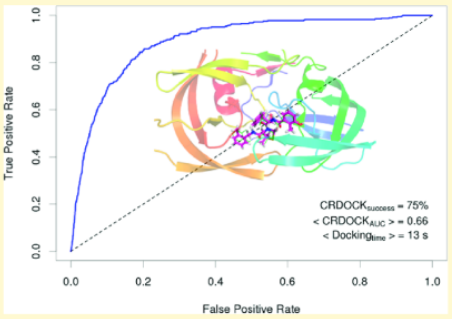
[§]SmartLigs Bioinformática S.L., Fundación Parque Científico de Madrid, c/Faraday 7, Campus de Cantoblanco UAM, E-28049 Madrid, Spain

[∥]Instituto de Biomedicina de Valencia (IBV-CSIC), c/Jaime Roig 11, E-46010 Valencia, Spain

[⊥]Instituto de Química Médica (CSIC), c/Juan de la Cierva 3, E-28006 Madrid, Spain

Supporting Information

ABSTRACT: An ultrafast docking and virtual screening program, CRDOCK, is presented that contains (1) a search engine that can use a variety of sampling methods and an initial energy evaluation function, (2) several energy minimization algorithms for fine tuning the binding poses, and (3) different scoring functions. This modularity ensures the easy configuration of custom-made protocols that can be optimized depending on the problem in hand. CRDOCK employs a precomputed library of ligand conformations that are initially generated from one-dimensional SMILES strings. Testing CRDOCK on two widely used benchmarks, the ASTEX diverse set and the Directory of Useful Decoys, yielded a success rate of ∼75% in pose prediction and an average AUC of 0.66. A typical ligand can be docked, on average, in just ∼13 s. Extension to a representative group of pharmacologically relevant G protein-coupled receptors that have been recently cocrystallized with some selective ligands allowed us to demonstrate the utility of this tool and also highlight some current limitations. CRDOCK is now included within VSDMIP, our integrated platform for drug discovery.



1. INTRODUCTION

Docking and virtual screening (VS) strategies have acquired a relevant role in modern drug discovery since the pioneering work of Kuntz et al.[1] back in the early 1980s. However, and despite many advances carried out in the field during the past decade, this methodology is still far from perfect.[2] To increase its usefulness, more accurate methods are needed that can not only predict the native pose of a ligand within a protein in a crystallographic structure at the top of the list of possible solutions, as done in docking studies, but also discriminate true binders from a pool of decoys, as done in VS. Moreover, a modern docking tool needs to be fast because the number of molecules in currently used chemical libraries is well above 10^6. It is clear then that different objectives are pursued in docking and VS. In the former, native pose prediction is the main goal; in the latter, as long as true binders are separated from decoys, less accurate binding poses can be tolerated. In an ideal case, however, both criteria should be met because success in VS for the wrong reasons is unlikely to be reproducible and does not contribute to advancing the field.[3] More commonly, the goodness of fit between the ligand and the receptor is evaluated by means of an energy function composed of different terms that attempt to account for the forces driving the binding event. Although the underlying physical laws describing the binding process are well understood, accuracy and computational resources (mainly time) evolve in opposite directions, and fine tuning the appropriate balance between them is by no means an easy task. Therefore, accuracy is normally sacrificed for speed, especially in VS, and very often too simplistic scoring functions are employed.

Prior to scoring it is also necessary to sample the binding site of the receptor as exhaustively as possible. To this end and to save computer time, the space is usually discretized on a three-dimensional (3D) lattice and probe interaction energies at the grid points are calculated and stored.[4] Then, the molecule under study is translated and rotated at each lattice node along the three dimensions of the box, and interaction energies with the protein are estimated for each pose using the data stored on the grid points. Depending on the docking tool, the conformers can be generated within the binding site itself "on the fly" or created beforehand and stored to be reused again as many times as needed. The former method is more computer intensive, but as an advantage, it can generate strained

Received: April 19, 2012
Published: July 5, 2012

conformations that adapt better to the active site environment. If the latter method is employed, a collection of all allowed conformers is quickly generated only once following some predefined rules, but the drawback is that a relevant conformation can be missed. For example, AutoDock[5] and GOLD[6] produce conformers in situ, whereas Glide[7,8] and FRED[9] use a pregenerated database of conformers. Our original in-house docking tool, CDOCK,[10] belongs to this second category because it was developed bearing in mind its potential use in VS, where millions of small molecules, some of them with hundreds of possible conformers, are available for docking in different projects. CDOCK was implemented in our open-access Virtual Screening Data Management on an Integrated Platform (VSDMIP) which has been recently extended to cover not only receptor-based VS[11] but also ligand-based[12] and fragment-based VS.[13]

Among the challenges still to be faced, some are more technical and related to computational times and correct implementation of tools to configure distinct VS protocols while others have to do with explicit inclusion of entropy,[14] solvent effects,[15,16] and receptor flexibility,[17,18] which appear to be necessary to get more accurate estimates of binding free energies. Indeed, theoretical approaches that tackle these problems continue to be developed and improved but are usually impractical for large VS campaigns because of the huge number of compounds that need to be evaluated. These computational efforts can be highly reduced if a prioritized list of compounds is available to be passed on to the most demanding calculations. This should be, in fact, the final objective of a docking program: saving time without increasing the false positive and negative rates compared to random selection. Unfortunately, not many currently available tools meet this requirement.

Bearing these facts in mind and in an attempt to improve the sampling and scoring capabilities of CDOCK, we present here CRDOCK, an ultrafast ligand docking program that was tested on two widely used benchmarks: the ASTEX diverse set (ADS, for docking)[19] and the Directory of Useful Decoys (DUD, for VS).[20] The latter was subsequently expanded with a representative group of recently available and pharmacologically very relevant G protein-coupled receptors (GPCR) in complex with some selective ligands. Being aware of the importance of water-mediated interactions in receptor−ligand binding, a water selection algorithm was implemented that is based on interaction energy calculations on a 3D grid, as pioneered by Peter Goodford in his renowned GRID program.[4] We also detected, discussed, and in most cases solved some widely reported problems[19,21] involving several well-known ligand−receptor complexes.

2. METHODS

Our CRDOCK tool contains (1) a search engine that can use a variety of sampling algorithms and an initial energy evaluation function for placing the ligand in the binding site, (2) several energy minimization algorithms for fine tuning the binding poses, and (3) different scoring functions for ligand ranking. Different methods are available for each of these components, and they can be independently chosen and combined.

2.1. Ligand Preparation. The ligands present in the complexes studied were prepared in two different ways. To test the docking engine alone, their X-ray coordinates from the PDB files were extracted and the ligand-free protein was used as the target (rigid ligand docking). Protonation and tautomeric states for these ligands were assigned with Open Babel 2.3.0[22] assuming a pH of 7.0, and no further manipulation of the coordinates was performed so as to preserve the native bound conformations. As a more realistic alternative and to check the efficiency of our in-house conformer generator, we also started from scratch, in the absence of any information about the ligand's 3D structure (flexible ligand docking). To this end, each ligand from the previous step was converted into a 1D simplified molecular-input line-entry system (SMILES) string and then inserted into the VSDMIP database following our standard protocol: (a) automated conversion from SMILES to 3D MOL2 format using CORINA,[23] (b) atomic charge calculations with MOPAC[24] (AM1 ESP method) on every single structure provided by CORINA, (c) atom-type assignment according to the AMBER force field,[25] and (d) conformer generation using ALFA.[26]

2.2. CRDOCK Constituent Parts. The three main components of the CRDOCK tool are depicted in Figure 1.

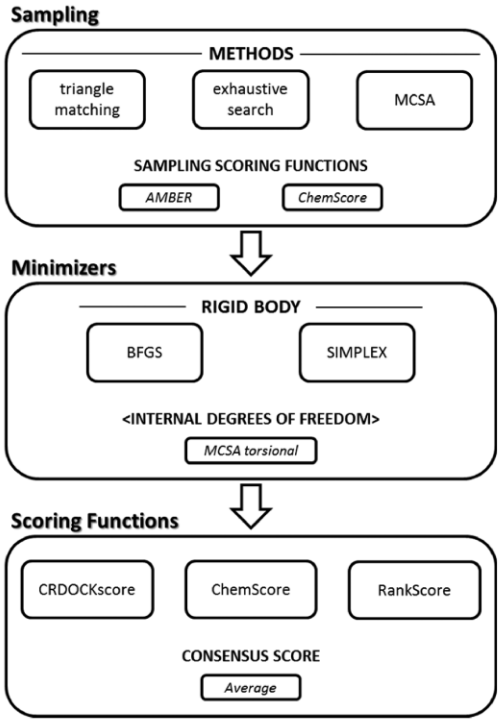


Figure 1. Graphical overview of CRDOCK components and workflow. Method enclosed in angle brackets is optional.

The different methods integrated in each step can be independently selected and later combined to configure different custom-made workflows. This is of special interest in those cases where a researcher faces computational restrictions.

2.2.1. Sampling. CRDOCK implements three sampling strategies: (a) triangle matching, (b) exhaustive search, and (c) Monte Carlo simulated annealing (MCSA).

a. Triangle Matching. CRDOCK determines the interaction points for a given ligand conformation using the functional groups present in its structure. All possible combinations of 3 different interaction points (for each conformer) are generated

applying a cutoff to avoid very short edges (4.5 Å by default). These are the ligand interaction triangles. The same combination is performed for the receptor interaction points (see receptor active site analysis below) to obtain the receptor interaction triangles. The program then looks for the best superimpositions between the receptor's and the ligand's interaction triangles, evaluating each one with an AMBER-like (12−6 Lennard−Jones potential with an electrostatic term modeled with a sigmoidal dielectric screening function) or the ChemScore[27] empirical scoring function. If none of the ligand conformers can be used to build triangles, a lower cutoff for the edges is employed (2.5 Å). Finally, if no triangle can be built with this reduced distance, the exhaustive search or MCSA algorithms (see below) will be used instead.

b. Exhaustive Systematic Search. In this case each ligand conformer is translated over each single grid point and rotated in all directions with steps of 30° on each axis. The interaction energy for each generated pose is evaluated with the energy functions described in point a above. Because this search is quite time consuming, it is performed only if the number of conformers is ≤5. For the remaining cases the MCSA method described below is employed.

c. Monte Carlo Simulated Annealing. Random translations and rotations of the ligand are generated to determine a new pose from the last accepted pose (the first pose is generated randomly). The new pose is accepted or rejected depending on a probability managed by the temperature (Metropolis criterion). By default and on the basis of results from internal tests, 23 temperature rounds are performed with a maximum number of generated poses per round of 725 000 to avoid redundancy. The algorithm starts at 773 K, and this temperature is scaled down at each round by 20%. The probability of change for each parameter defining a pose was set to 0.8.

Regardless of the sampling method, the best 512 poses per conformer are saved in a stack up to a maximum of 5000 per molecule. The final result is a collection of the best ligand poses (5000 by default) which can be passed on to the next step to be refined by energy minimization. In order to promote diversity, it is possible to apply a root-mean-square deviation (rmsd)-based filter to check whether or not a new pose is added to the list.

2.2.2. Minimization. Prior to scoring, the above selected poses can be refined either by rigid-body energy minimization (translations and rotations) or by changing the internal degrees of freedom (torsional angles) to fine tune their fit within the receptor binding site. This process is carried out expediently using precalculated AMBER-like potential interaction energies stored in a 3D grid.[28,29] Three algorithms have been implemented: (a) Broyden−Fletcher−Goldfarb−Shanno (BFGS, for rigid-body minimizations), a deterministic method that belongs to the family of quasi-Newton methods; (b) Amoeba or downhill SIMPLEX (for rigid-body minimizations), the stochastic algorithm from Nelder and Mead;[30] and (c) MCSA torsional (MCSAtor, for internal degrees of freedom [torsions] minimizations), a stochastic method analogous to the described MCSA algorithm but that only optimizes molecular torsions. By default, 200 steps of BFGS rigid-body minimization are performed.

2.2.3. Scoring Functions. The final ranking of the resulting poses from the previous step can be evaluated with either a single scoring function or a combination of several of them ("consensus scoring"). The available options are (a) CRDOCKscore, (b) ChemScore, and (c) RankScore.

a. CRDOCKscore. This is a modified version of GlideScore,[31] the scoring function implemented in program Glide (eq 1), that combines van der Waals (E_{vdW}) and electrostatic (E_{qq}) energy terms from the AMBER force field with lipophilic (E_{lipo}) and hydrogen-bonding (E_{hb}) terms from ChemScore.[27] The AMBER terms are scaled by weighting factors α and β for E_{vdW} and E_{qq}, respectively, as defined in GlideScore.

$$\text{CRDOCKscore} = \alpha E_{vdw} + \beta E_{qq} + E_{lipo} + E_{hb} \quad (1)$$

where α = 0.065 and β = 0.130. No additional modifications were made to the AMBER or ChemScore lipophilic and hydrogen-bonding terms.

b. ChemScore. ChemScore is our implementation of the well-known ChemScore[27] scoring function.

c. RankScore. RankScore is a statistical potential scoring function specifically derived for VS classification from known sets of compounds (DUD).[32]

2.2.4. Water Selection Algorithm. Very often water-mediated interactions between the ligand and the receptor binding site are key for accurate fitting. We employed a modified version of cGRILL (see receptor active site analysis) to generate water affinity maps using a probe representing a water molecule that can act as a hydrogen-bond acceptor and donor. For this purpose we employed the concept of *extended atom* using an AMBER-like energy function for an oxygen atom endowed with a van der Waals term, a partial charge of −0.12 au and a hydrogen-bond block function based on an ideal H-acceptor distance of 1.8 Å and a donor−H-acceptor angle of 180°. Then the program clusters similar interaction areas using an energy cutoff which is one-half of the maximum value of the scoring function.

2.3. Benchmarks. Three different test sets were used: (a) ADS for pose prediction, (b) DUD for VS, and (c) GPCR for pose prediction and VS. The standard criterion to validate the ability to predict the native pose was the heavy atoms rmsd between the docking solution and the native conformation for each ligand in the crystal structure. In common with other similar studies, we chose an rmsd value of 2 Å as the upper limit for a solution to be considered correct. VS performance was assessed by means of the area under the curve (AUC) of the generated receiver operating characteristics (ROC) plots.[33]

All receptors (except otherwise stated) were prepared for both docking and VS following the same protocol, namely, for each one, all species other than the protein itself, cofactors, and metals were removed. Hydrogen atoms were added using the pdb2pqr[34] tool and adapted to the AMBER 03 force field using GROMACS v.4.5.3. One thousand steps of steepest descent were followed by 2000 steps of Polak−Ribiere conjugate gradient energy minimization where only hydrogen atoms were allowed to move. No energy minimization was performed for DUD targets so that our results could be compared with those published in other DUD-related publications. The cubic grid for cGRILL calculations was defined as the space delimited by the axis-parallel box containing the cocrystallized ligand, augmented by 5 Å in each axis direction.

a. ASTEX Diverse Set (ADS). The ADS is composed of 85 protein−ligand complexes that can be downloaded from the Protein Data Bank (PDB).

b. Directory of Useful Decoys (DUD). The DUD is composed of 40 different targets with known 3D structures and a set of true/fake binders for each target.

c. G Protein-Coupled Receptors (GPCR) Set. We enlarged the original DUD by including such pharmacologically relevant targets from the GPCR family as the adrenergic β_2, the dopaminergic D_3, the muscarinic M_2, the histamine H_1, and the opioid μ receptors, all of which have been cocrystallized in the presence of antagonists or inverse agonists at a resolution ≤ 3.1 Å. For both self-docking and VS tests we used the ligand-bound protein structures found in PDB entries 2RH1 (β_2), 3PBL (D_3), 3UON (M_2), 3RZE (H_1), and 4DKL (μ). In the latter case, the covalent bond between the morphinan ligand and Lys233 was broken and an alternate nonclashing rotamer was selected for Lys233 using PyMOL.[35] SMILES strings for known true ligands for these targets (Supporting Information, Table S1) were taken from the DrugBank database.[36] To select a suitable set of decoys, we followed a similar procedure to that reported in the original DUD description: (a) the clean drug-like set of small molecules was downloaded from ZINC[37] (9 542 593 SMILES strings); (b) Molecular ACCess System (MACCS) fingerprints were calculated for each SMILES string using OpenBabel;[22] (c) a Tanimoto cutoff of 0.4 was used as a filter to select the most topologically dissimilar compounds as compared to the known ligands (415 636); (d) Qikprop[38] was used to calculate physicochemical properties for each compound and select those most similar to the true ligands; and (e) 30 decoys per ligand were saved for the VS experiments.

2.4. Receptor Ligand-Binding Pocket Analysis. Given a suitably prepared receptor structure, the next steps are to energetically characterize the active site and to determine its main interaction points.

a. Energetic Characterization. This is performed with the cGRILL program, an improved version of our CGRID code[10] that relies on interaction energy calculations on 3D grids as pioneered by Goodford in his well-known GRID program.[4] cGRILL uses the AMBER 12−6 Lennard−Jones term for C, N O, H, S, and P probe atoms and an electrostatic term modeled with a sigmoidal dielectric screening function, together with other terms from ChemScore (lipophilic, H-bond acceptor, H-bond donor, "mixture", metal, and clash). Finally, a grid containing clash-free points was added for accelerating the sampling process during docking. All grid maps can be inspected using PyMOL molecular visualization software.[35] Residues, cofactors, and other species are parametrized automatically using an AMBER force field-like atom-typing scheme.

b. Interaction Points. The best interaction areas ("hot spots") are mapped by sampling the active site with different molecular probes, an idea already introduced in other docking programs.[39] The calculations are similar to those in cGRILL for AMBER grids but using polyatomic probes instead, which are allowed to rotate at each grid point with a step size of 180° in each axis. The molecular probes are CH_4 to detect lipophilic regions and NH and CO to detect H-bond-accepting and -donating partners, respectively. The best result for any of the probes at each grid point is selected. The generated set is postprocessed to avoid redundancy. For a given area, only the best probe in a radius of 2 Å for lipophilic or 1.5 Å for hydrogen-bond probes is kept. Then, these nonredundant probes are rescored by summing up their neighbors' energetic scores.[39] The best probe with the highest new score is selected as the best point in the active site, and all the surrounding points with a distance less than 4 Å are also selected. The process continues adding points within 4 Å from the selected ones until no more points fulfilling the distance restraints are available. The selected docked probes are exported as a PDB file.

2.5. Alternative Docking Protocols. Two different alternatives were explored.

(a) Alternative 1. Our original CDOCK code with default parameters: MCSA/exhaustive sampling, SIMPLEX refinement, and the scoring function as the sum of van der Waals and electrostatic terms from the AMBER force field, the electrostatic contributions to ligand and receptor desolvation, the nonelectrostatic part of the desolvation modeled as a linear relationship with the solvent accessible surface area lost once the complex has been formed, and an explicit term to account for hydrogen-bonding interactions. In the preparation of the receptor, water molecules involved in the binding event are always kept and full complex relaxation is accomplished using energy minimization.

(b) Alternative 2. A different CRDOCK configuration consisting of (1) default sampling (triangle matching or MCSA), (2) BFGS and MCSAtor for fine tuning, and (3) nonbonding energy terms (E_{vdW} and E_{qq}) from the AMBER force field for pose scoring. For the VS experiments, ChemScore, AMBER, and RankScore were used as the scoring functions as well as a simple average of the three ("consensus scoring").

3. RESULTS AND DISCUSSION

Different CRDOCK configurations were tested for both data sets. After analyzing the results (Table 1), the combination that yielded the best performance was triangle matching and AMBER energy evaluation, BFGS as the rigid body minimizer without torsional optimization, and the CRDOCK scoring function. The results shown in Table 2 were obtained with this CRDOCK configuration.

Table 1. Summary of the Results from Alternative Protocol Using the ADS and the DUD

protocol	no. complexes with rmsd ≤ 2.0 Å (ADS)	average AUC (DUD)
CRDOCK	62	0.66
alternative 1 (CDOCK)	64[e]	0.60
alternative 2 FF[a]	59[f]	0.56
alternative 2 CS[b]	59[f]	0.58
alternative 2 RS[c]	59[f]	0.57
alternative 2[d]	59[f]	0.58

[a]AMBER force field scoring function. [b]ChemScore. [c]RankScore. [d]Consensus scoring = average between AMBER, Chemscore, and Rankscore scoring functions. [e]See methods. [f]Selected poses are always the same, and the difference is due to the final scoring function applied.

3.1. Docking and Scoring: Pose Prediction and Errors Found. Although with some controversy,[40] the rmsd between a docking solution and the crystallographic coordinates is still the most widely accepted criterion as a metric of success. A docking solution with an rmsd value ≤ 2.0 Å is regarded as a correct pose. Considering the receptor as a rigid entity, two different docking experiments were conducted: (a) rigid ligand docking, where the ligand conformation is taken directly from the structure of the complex, and (b) flexible ligand docking, where

Table 2. Summary of the VS Statistics Related to the AUC Values Grouped by Protein Families

target[a] family	average	SD[b]	median	max	min
global	0.66	0.15	0.69	0.97	0.32
kinases[c]	0.63	0.11	0.63	0.77	0.47
serineproteases[d]	0.78	0.06	0.76	0.84	0.74
NHR[e]	0.71	0.19	0.74	0.97	0.42
metalloenzymes[f]	0.61	0.15	0.63	0.76	0.41
folateenzymes[g]	0.83	0.15	0.83	0.94	0.72
rest	0.61	0.26	0.59	0.80	0.32

[a]ACE, angiotensin-converting enzyme; AChE, acetylcholinesterase; ADA, adenosine deaminase; ALR2, aldose reductase; AmpC, AmpC β-lactamase; AR, androgen receptor; CDK2, cyclin-dependent kinase 2; COMT, catechol *O*-methyltransferase; COX-1, cyclooxygenase-1; COX-2, cyclooxygenase-2; DHFR, dihydrofolate reductase; EGFR, epidermal growth factor receptor; ER_{ago}, estrogen receptor (agonist-bound conformation); ER_{antago}, estrogen receptor (antagonist-bound conformation); FGFR1, fibroblast growth factor receptor kinase; FXa, factor Xa; GART, glycinamide ribonucleotide transformylase; GPβ, glycogen phosphorylase β; GR, glucocorticoid receptor; HIVPR, HIV protease; HIVRT, HIV reverse transcriptase; HMGR, hydroxymethyl-glutaryl-CoA reductase; HSP90, human heat shock protein 90; INHA, enoyl ACP reductase; MR, mineralocorticoid receptor; NA, neuraminidase; P38 MAP, P38 mitogen-activated protein; PARP, poly(ADP-ribose) polymerase; PDE5, phosphodiesterase 5; PDGFRB, platelet-derived growth factor receptor kinase; PNP, purine nucleoside phosphorylase; PPARγ, peroxisome proliferator activated receptor γ; PR, progesterone receptor; RXRα, retinoic X receptor α; SAHH, *S*-adenosyl-homocysteine hydrolase; SRC, tyrosine kinase SRC; Thr, thrombin; TK, thymidine kinase; VEGFR2, vascular endotelial growth factor. receptor; [b]Standard deviation. [c]CDK2, EGFR, FGFR1, HSP90, P38 MAP, PDGFRB, SRC, TK, and VEGFR2. [d]FXa, Thr, and trypsin. [e]AR, ER_{ago}, ER_{antago}, GR, MR, PPARγ, PR, and RXRα. [f]ACE, ADA, COMT, and PDE5. [g]DHFR and GART.

a precalculated set of conformers is generated from scratch. In the former we challenge, on one hand, whether the sampling algorithms are able to generate the correct pose and, on the other hand, whether the scoring function is able to recognize it as its best solution. In the latter we also test the conformer generator. Our results (Figure 2) indicate that CRDOCK correctly identifies 93% (79 out of 85) of the native poses in the rigid docking experiment. However, this percentage falls to 73% (62 out of 85) when the ligand coordinates are generated from the SMILES strings using CORINA and ALFA.[26] Both results are comparable to those obtained with the most widely used docking programs. For example, in the original ADS study the authors described a very similar performance for GOLD (around 70−80% of complexes with an rmsd $\leq$ 2.0 A) and almost the same decrease in performance when the ligand 3D structure was generated from scratch instead of using the crystallographic pose.[19] Li et al. used a set of 195 diverse high-resolution ligand−protein complexes to compare Glide, GOLD, LigandFit, and Surflex docking programs and obtained a variable range from 60% to 80% in most cases.[41] Cross et al., using a subset of high-resolution structures from the CCCD/Astex set, compared GLIDE, ICM, PhDock, FlexX, and Surflex docking tools. The 70−90% rate of successfully docked poses when the native ligand was reduced to 50−77% when the ligand structure was generated from scratch using CORINA.[42] More recently and using ADS, performances of 60−80% for Surflex-Dock[43] and ~74% for LeadFinder[44] were reported. All benchmarks confirm than the results currently obtained with CRDOCK are comparable to those obtained with other modern docking programs. Next, we looked for the reasons for failure.

Common Errors. Some errors that we encountered in flexible ligand docking appear to be commonly reported by others[19] using different docking engines and scoring functions.

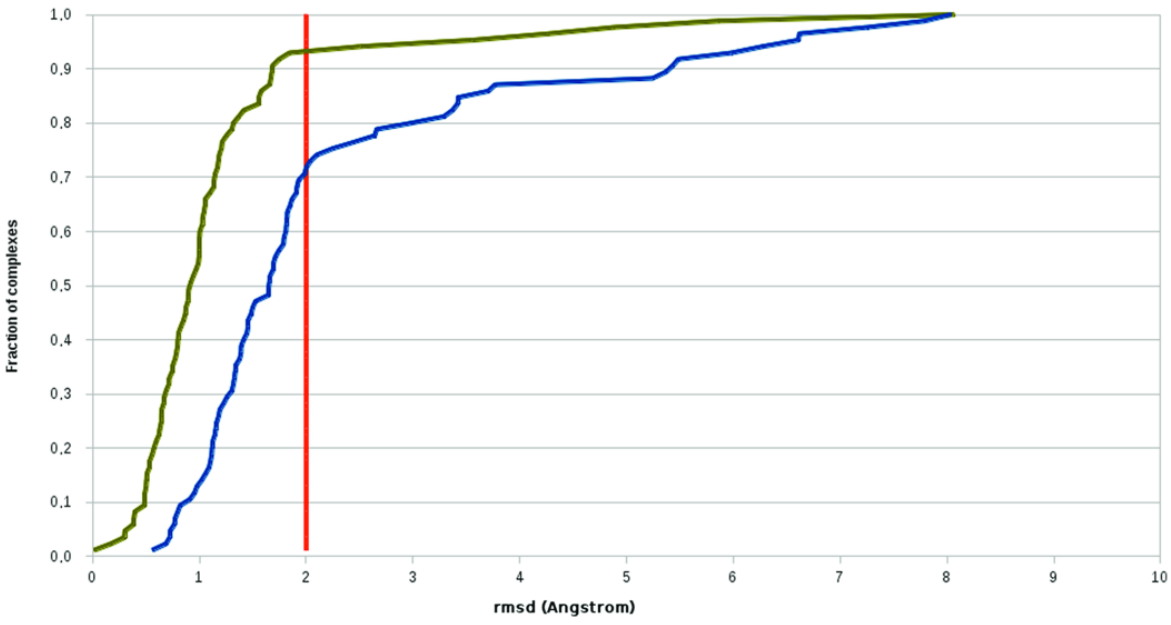


Figure 2. Pose prediction performance obtained using CRDOCK in rigid (green) and flexible (blue) ligand docking. Red line indicates a success threshold of rmsd = 2 Å.

The ligands in PDB entries 1JJE, 1SJ0, 1YVF, and 1TZ8 belong to this category (Supporting Information, Figure S1). The ligand in complex 1JJE displays an almost planar and quasi-symmetrical conformation with a central diacetal moiety chelating a zinc ion. Due to these properties it is not unreasonable that the docking program, while being able to capture the native interactions, selects a pose that is rotated by 180° compared to the X-ray structure. The same is observed for the symmetric ligand in the 1TZ8 complex. The best docking solution reproduces all of the contacts present in the X-ray structure, but its rmsd (2.64 Å) is (only apparently) beyond the success cutoff. The native pose for the ligand in the 1YVF complex has a carboxylic group directly exposed to the solvent whose location is strongly penalized by the scoring function. Therefore, the native pose is not promoted to the top of the ranking list. Finally, the ligand in the 1SJ0 complex presents, in its central ring, a large substituent in the axial position and a small substituent in the equatorial position. This peculiar arrangement is not considered stable enough to be selected by CORINA as a representative conformation, and therefore, the docking algorithm fails in reproducing the native pose.

Another common error is the sulfonamide ligand in 1JD0, which was not docked correctly into human carbonic anhydrase XII using the rigid docking protocol (rmsd = 5.7 Å) but, rather surprisingly, was successfully docked (rmsd = 1.85 Å) when starting from scratch.

Ligand Conformational Sampling Errors. These errors are related to our ligand preparation step. In some cases the generated set of conformers does not properly cover the conformational space of a ligand. In the present study, the vast majority of the problematic cases are those in which the bound ligand in the experimental structure exhibits torsional angles whose values are away from those considered as ideal. These torsions are, to some extent, forced by the binding site environment. Since ideal angles are taken as rules in our conformer generator code ALFA, these particular conformations will not be present in the conformer population. Within the ADS, the ligands found in PDB entries 1Q1G, 1R58, 1UML, 1UNL, and 1UOU fall into this category.

To test whether or not this was, in fact, the source of error, these ligands were parametrized (GAFF, General Amber force field), immersed in cubic boxes of TIP3P water molecules, energy minimized (2000 steps of steepest descent followed by 2000 additional steps of Polak−Ribiere conjugate gradient), and simulated (constant temperature [300 K] and pressure [1 atm]) for 15 ns. Torsional angles were monitored over the whole trajectory and compared to those found in the crystal. In all five cases the relative population of conformers whose torsional angles were within 5° of the crystallographic structure was below 7%. This means that the native conformation is rarely sampled, and therefore, algorithms based on rules are very unlikely to generate near-native conformations.

Errors Due to Missing Water-Mediated Interactions. There are five cases of failure in which water molecules crucially mediate ligand−protein interactions (water-mediated bridges). As the usual protocol consists of eliminating, among other species, all water molecules present in a binding site, docking algorithms are unable to reproduce the native pose. This problem occurs in PDB entries 1GM8, 1G9V, 1GPK, 1HVY, and 1XM6 (Supporting Information, Figure S1) and was also found in the rigid ligand docking experiments. The ligand in complex 1GM8 has a β-lactam ring that interacts with residue Ser386 through a water-mediated bridge (WAT2460) and a benzylic amide buried at the bottom of the binding pocket. CRDOCK correctly reproduces the X-ray position of the latter moiety, while it places the β-lactam ring in an alternate location (rmsd = 3.37 Å). In the 1G9V complex, WAT916 mediates the interaction between Asn308 and the nitrogen atom of the ligand's amide group whereas WAT927 and WAT983 interact with the ligand's carboxylate. Both functional groups are wrongly positioned by CRDOCK in the absence of the water molecules (rmsd = 3.77 Å). The ligand in the 1GPK complex has a charged amino group interacting with a water molecule (WAT2529). When the water molecule is not considered as part of the binding site, the amino group is attracted by Glu199 and an alternative docking pose is favored (rmsd = 3.71 Å). The ligand in complex 1HVY has two different sources of error. First, the conformational sampling of the ligand fails because the closest conformer found is 1.47 Å away from the X-ray structure; second, the absence of two water molecules (WAT477 and WAT1017) precludes the correct orientation of the two carboxylic moieties at one end of the molecule (rmsd = 3.36). Finally, the ligand in the 1XM6 complex is unable to correctly orientate the oxazolidinone ring if WAT1009, which chelates the Zn^{2+} ion, is not present in the binding site. In its absence, the ligand's carbonyl group is strongly attracted by the Zn^{2+} ion (rmsd = 2.45).

The "missing water" problem could be easily solved in some cases using cGRILL to calculate the water affinity map within the binding site and place the relevant water molecule(s). These were correctly identified for both 1G9V and 1XM6, and upon incorporation into the protein structure, the rmsd of the ligands was notably reduced relative to the "dry" docking solution (from 3.77 to 1.51 Å and from 2.45 to 1.90 Å for 1G9V and 1XM6, respectively). For the remaining cases adding the water molecules helped only in part. In the 1HVY complex the ligand's carboxylate groups were correctly positioned upon addition of the water molecules, but the conformational problem still persisted (rmsd = 2.75 Å). Proper docking of the ligand in the 1GM8 complex improved significantly following incorporation of the water molecules; most of the contacts with the target were reproduced, but the rmsd was still 2.67 Å. On the other hand, for the same ligand in the "rigid" approach the rmsd decreased from 4.27 to 0.61 Å. Finally, for the 1GPK complex, addition of the relevant water molecules resulted in a pose with rmsd = 1.39 Å, although this solution was the second best on the ranking list.

Other Errors. Some compounds were classified as misdocked because of an rmsd slightly above 2.0 Å despite the fact that the top scoring pose faithfully reproduced the main interaction points that are observed within the binding site in the X-ray crystal structure. Five complexes belong to this category of "soft errors": 1HP0 (rmsd = 2.57 Å), 1N2V (rmsd = 2.86 Å), 1MMV (rmsd = 2.84 Å), 1XOQ (rmsd = 2.10 Å), and 2BM2 (rmsd = 2.25 Å). Finally, the ligands in complexes 1IG3, 1OQ5, 1P62, and 2BSM were not correctly positioned and no clear reason could be found to account for the deviations, which are presumably due to several combined factors such as insufficient or deficient ligand conformational sampling and/or inaccuracies in the scoring function.

3.2. Virtual Screening: Discriminating True Binders from Decoys. VS calculations were performed with the 40 targets and their associated sets of true binders and decoys from DUD. Global as well as family-wise statistics (AUCs from the ROC plots) were compiled and are summarized in Table 2. The average AUC values are comparable to those already

reported,[45] and there is a clear variability depending on the target. The 5 top scores are obtained for RXRα (AUC = 0.97), DHFR (AUC = 0.94), ER_{ago} (AUC = 0.86), trypsin (AUC = 0.84), and COX2 (AUC = 0.80). These targets are related neither functionally nor structurally, except for ER_{ago} and RXRα which belong to the family of nuclear hormone receptors (NHR). The same lack of obvious connection applies to the worst scoring targets, with the exception of the two NHR members PR and PPARγ: TK (AUC = 0.47), PR (AUC = 0.45), PPARγ (AUC = 0.42), ACE (AUC = 0.41), and AmpC (AUC = 0.32). The trend that CRDOCK performs better than average on folate enzymes and serine proteases has been observed with other docking programs.[45] In the case of kinases our method outperforms other docking tools reporting on the same data set[45] even though the average AUC value for this family is still below the global AUC average.

We took a closer view at the ROC curves (Supporting Information, Figure S2) to shed more light into those targets for which VS performance was worse than random: AChE (AUC = 0.49), TK (AUC = 0.47), PR (AUC = 0.47), ACE (AUC = 0.41), InhA (AUC = 0.48), and AmpC (AUC = 0.32). AmpC appears to be one of most problematic targets for all docking programs.[42] AChE[46] and TK[47] are two well-known enzymes in which the flexibility of some residues within the active site plays a crucial role in ligand recognition. As our docking tool does not include receptor flexibility, it is not entirely surprising that the VS protocol that we used fails in these cases, as do many other programs.[42] Although the global AUC values were below random for PR, ACE, and InhA, early enrichments (as detected in other studies)[42] were clearly apparent because 3/5, 4/9, and 13/17 true binders, respectively, were present on the top 0.5% of the rank-ordered list.

Different docking programs using DUD as a test set afforded values similar to those reported here for CRDOCK. Cross et al.[42] found average AUC values of 0.55 (DOCK), 0.59 (PhDock), 0.61 (FlexX), 0.63 (ICM), 0.66 (Surflex), and 0.72 (Glide). Finally, Marco et al.[48] reported a median AUC of 0.69 (very similar to CRDOCK) and Novikov et al.[44] an average AUC around 0.70.

3.3. GPCR: Pose Prediction and Virtual Screening. The original ADS was supplemented with five GPCR complexes, and self-docking experiments were carried out as described in the pose prediction section. The ligands in PDB entries 3PBL, 3UON, 2RH1, and 3RZE were correctly positioned within the binding site (rmsd $\leq$ 2.0 Å) in both the rigid and the flexible ligand docking tests. In the case of the μ opioid receptor, however, the first solution had rmsd = 2.44 Å, but it has to be borne in mind that in the crystal structure (PDB entry 4DKL) the funaltrexamine ligand is covalently bonded to Lys233. In the CRDOCK solution (Supporting Information, Figure S3) the main deviation lies precisely on the flexible chain that is covalently bonded to the amino group of Lys233 in the experimental structure, whereas the morphinane core reproduced the native pose accurately (rmsd = 1.27 Å). Therefore, although the overall rmsd is above the canonical cutoff value of 2.0 Å, the docking error can be considered "soft" in light of the conserved important interactions.

To explore the ability of CRDOCK to predict selectivity, we performed a simple cross-docking VS experiment using these five GPCR complexes, that is, every ligand was docked into every receptor to check whether the best score was assigned to the true ligand−receptor couple. In addition, by superimposing the seven transmembrane helices of all the GPCR studied we were able to calculate and compare the rmsd for all docking poses (Table 3). Without exception, the lowest rmsd was found for the native ligand in its own receptor. In addition, a very low rmsd at the M_2 receptor was found for doxepin, an H_1 receptor antagonist which is also known to bind with high affinity to muscarinic,[49] α_1-adrenergic,[50] and some mosquito dopamine[51] receptors. Regarding the ranking, the native ligands received the top scores at the β_2, M_2, and μ opioid receptors but those of the D_3 and H_1 receptors got the second best. At the D_3 receptor it was doxepin that appeared in the first position, and at H_1 it was not doxepin, as expected, but 3-quinuclidinyl benzilate, the prototypical anticholinergic agent with high selectivity for M_2 receptors. These findings raised a note of caution and prompted us to perform further studies.

Table 3. RMSD Values (in Angstroms, top) and Scores (kcal mol^{-1}, bottom) for Cross-Docking Studies on GPCRs

	ligands				
	β_2	D_3	H_1	M_2	μ
GPCR[a]					
β_2	**1.10**	3.81	5.67	4.58	5.42
D_3	3.80	**2.00**	5.10	6.80	7.18
H_1	4.29	3.98	**1.61**	4.48	8.11
M_2	4.67	4.60	**2.16**	**1.15**	5.93
μ	6.03	5.21	8.51	4.72	**1.27**
GPCR[b]					
β_2	**−101.3**	−86.1	−85.6	−82.8	−46.2
D_3	−83.9	**−94.0**	*−98.2*	−91.0	−73.7
H_1	−111.2	−97.3	**−117.5**	*−123.0*	−107.7
M_2	−87.8	−83.5	−104.7	**−115.4**	−57.2
μ	−61.6	−77.5	−72.7	−88.0	**−92.1**

[a]Numbers in bold along the diagonal highlight the lowest rmsd values for the ligand bound to its cognate receptor and, in addition, for the H_1 ligand (doxepin) bound to the M_2 receptor. [b]Numbers in bold along the diagonal highlight the binding energies for the cognate ligands, which are sometimes less favorable (in the same row) than for a noncognate ligand (values in italics).

When all true (as annotated in the DrugBank[36]) GPCR ligands studied were merged into a single set and used in an expanded VS experiment against the five GPCR, the AUC values from the obtained ROC curves were 0.50 (D_3), 0.53 (β_2), 0.56 (M_2), 0.71 (μ), and 0.61 (H_1). More informative, however, is the number of true binders that are recovered for each target at the top 10 of the ordered list: 6 are correctly identified for the β_2, D_3, and H_1 receptors, 5 for the M_2 receptor, and 2 for the μ opioid receptor.

Finally, when the set of ligands for each GPCR consisted of 30 decoys per true binder, the AUC from the resulting ROC curves were 0.50 (D_3), 0.56 (β_2), 0.64 (M_2), 0.80 (μ), and 0.82 (H_1). The average value of 0.67 compares very well with the findings reported above for the DUD. Nonetheless, it has to be noted that the number of true binders for each GPCR varied greatly, from 7 for the opioid receptor to 74 for the H_1 receptor (Supporting Information, Tables S1−S5), and this wide difference can be largely responsible for the diversity of the outcome. After completion of this work, we became aware of a recent compilation of 147 GPCR targets and a ligand library (agonists + antagonists) that included 39 decoy molecules for each true binder.[52] Application of the CRDOCK VS protocol to our selected GPCRs and this alternative compound

collection resulted in a small improvement for the β_2 receptor (0.1 AUC units), roughly the same performance for the D_3 receptor, and a slight decrease for μ (0.2 AUC units), H_1 (0.2 AUC units), and M_2 (0.1 AUC units) receptors. These differences, which amount to an average decrease in AUC from 0.67 to 0.56, can be expected due to the distinct way decoys and true ligands were selected and also to the fact that all receptors used are in the antagonist-bound conformation.

Altogether, these results are encouraging but indicate that to study selectivity among related GPCR with an acceptable degree of accuracy further improvements in CRDOCK and receptor preparation (e.g., incorporation of bound water molecules)[52] will be necessary.

3.3. Alternative Docking Protocols with Flexible Ligands. Two other alternatives were tested: our old in-house docking engine CDOCK and a different CRDOCK configuration.

From the results obtained for VS using DUD, CRDOCK represents a significant improvement over the rest (Table 1), which confirms the importance of the hybrid scoring function. However, performance on an individual target is in some cases strongly dependent on the selected scoring function. For example, in the case of the PDB code 1XGJ (β-lactamase AmpC) the AUC for the hybrid scoring function is as low as 0.32, but this figure is increased to 0.71when a force-field-based scoring function is used. Therefore and in agreement with other studies, we believe that trying to develop a universal scoring function[53,54] can be a daunting task indeed. Instead, more satisfactory results for the problem in hand can be achieved if a tailor-made target-dependent scoring function is used.

3.4. Benchmarking. The average time to perform the docking of a single flexible ligand using the reported combination of pieces that yielded the best performance is ~10 s on a 64-bit 3.3 GHz Intel Core i5 processor. We observed that the triangle matching approach could be used for around 91% of the compounds in the database. For the remaining ligands either exhaustive search or MCSA was used, and this took ~26 s on average to complete, leaving the docking average time in ~13 s. Therefore, with a modest cluster of 100 processors it should be possible to screen more than one-half million compounds per day in a typical VS campaign.

4. CONCLUSIONS

We introduce CRDOCK, a new computational tool that performs reasonably well in both pose prediction (docking) and true binder discrimination (VS). CRDOCK has demonstrated its abilities on two widely used benchmarking tests: the ADS for pose prediction and the DUD for VS. The docking failures found, some in common with other published reports, were analyzed in detail, and several solutions were found. In addition, five representative ligand−GPCR complexes were studied, and the results were in line with those obtained from DUD. Self-docking was always satisfactory with rmsd values below 2.0 Å, cross-docking was correct in 3 out of 5 cases, and the VS results provided encouraging results that support previous evidence[55] suggesting the feasibility of carrying out successful VS campaigns on pharmacologically important GPCR.[52,56] It is expected that future work on the remaining caveats will enhance CRDOCK performance.

Besides its accuracy, an additional advantage of CRDOCK is its reduced computational cost, once the conformational library for the ligands has been generated. It should be noted that this process is done only once, and the resulting conformers can be employed in different VS campaigns, so that millions of compounds can be screened thereafter using a relatively modest computational infrastructure. CRDOCK is open source and can be downloaded free of charge to noncommercial parties following registration at the CBM Bioinformatics Unit's web page (http://ub.cbm.uam.es/).

■ ASSOCIATED CONTENT

Supporting Information

Figures containing a structural depiction of docking errors, ROC curves corresponding to those targets for which VS performance was worse than random, and X-ray and docking solution for the ligand present in PDB entry 4DKL; tables listing the β_2, D_3, H_1, M_2, and μ binders used in the VS of GPCRs as well as the AUC values per DUD target. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: amorreale@cbm.uam.es.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by grants from CICYT (SAF2009-13914-C02-02 to F.G.) and Comunidad Autónoma de Madrid (S-BIO-0214-2006 [BIPEDD] and S2010-BMD-2457 [BIPEDD2] to A.M. and F.G.). A.M. and J.K. acknowledge financial support from Fundación Severo Ochoa through the AMAROUTO program and Ministerio de Economía y Competitividad (BFU2011-24595), respectively. R.G.-R. enjoyed a MICINN contract from "Programa de Personal Técnico y de Apoyo 2008", and A.C. is the recipient of FPU grant AP2009-0203 from the Ministerio de Educación. We are grateful to OpenEye Scientific Software, Inc. for providing us with an academic license for their software. The technical support and advice from the Bioinformatics team at CBMSO is gratefully acknowledged.

■ ABBREVIATIONS

AUC, area under the curve; BFGS, Broyden−Fletcher−Goldfarb−Shanno; DUD, Directory of Useful Decoys; MCSA, Monte Carlo simulated annealing; rmsd, root-mean-square deviation; ROC, receiver operating characteristic; VSDMIP, Virtual Screening Data Management on an Integrated Platform

■ REFERENCES

(1) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

(2) Woltosz, W. S. If we designed airplanes like we design drugs. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 159−163.

(3) Kolb, P.; Irwin, J. J. Docking screens: right for the right reasons? *Curr. Top. Med. Chem.* **2009**, *9*, 755−770.

(4) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(5) Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins* **1990**, *8*, 195−202.

(6) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.
(7) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.
(8) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(9) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68*, 76−90.
(10) Pérez, C.; Ortiz, A. R. Evaluation of docking functions for protein-ligand docking. *J. Med. Chem.* **2001**, *44*, 3768−3785.
(11) Gil-Redondo, R.; Estrada, J.; Morreale, A.; Herranz, F.; Sancho, J.; Ortiz, A. R. VSDMIP: virtual screening data management on an integrated platform. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 171−184.
(12) Cabrera, Á. C.; Gil-Redondo, R.; Perona, A.; Gago, F.; Morreale, A. VSDMIP 1.5: an automated structure-and ligand-based virtual screening platform with a PyMOL graphical user interface. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 813−824.
(13) Cortes-Cabrera, A.; Gago, F.; Morreale, A. A reverse combination of structure-based and ligand-based strategies for virtual screening. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 319−327.
(14) Marshall, G. R. Limiting assumptions in structure-based design: binding entropy. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 3−8.
(15) Yuriev, E.; Agostino, M.; Ramsland, P. A. Challenges and advances in computational docking: 2009 in review. *J. Mol. Recognit.* **2011**, *24*, 149−164.
(16) Mysinger, M. M.; Shoichet, B. K. Rapid context-dependent ligand desolvation in molecular docking. *J. Chem. Inf. Model.* **2010**, *50*, 1561−1573.
(17) Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M.; Orozco, M.; Pertinhez, T. A.; Rizzi, M.; Sotriffer, C. A. Target flexibility: an emerging consideration in drug discovery and design. *J. Med. Chem.* **2008**, *51*, 6237−6255.
(18) Kokh, D. B.; Wade, R. C.; Wenzel, W. Receptor flexibility in small-molecule docking calculations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*; Wiley: New York, 2011; Vol. *1*, pp 298−314.
(19) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−741.
(20) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(21) Liebeschuetz, J.; Cole, J.; Korb, O. Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test. *J. Comput.-Aided Mol. Des.* **2012**, Epub ahead of print.
(22) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 1−14.
(23) *Corina Molecular Networks*; GmbH Computerchemie Langemarckplatz 1, E., Germany, 2000.
(24) Stewart, J. J. MOPAC: a semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−105.
(25) Case, D. A.; Darden, T. A.; Cheatham, I., T.E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlam, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER 8*; University of San Francisco: San Francisco, 2004.
(26) Gil-Redondo, R. Master Thesis. UNED, Madrid, 2006.
(27) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.
(28) Gschwend, D. A.; Kuntz, I. D. Orientational sampling and rigid-body minimization in molecular docking revisited: on-the-fly optimization and degeneracy removal. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 123−132.
(29) Wang, J.; Kollman, P. A.; Kuntz, I. D. Flexible ligand docking: a multistep strategy approach. *Proteins* **1999**, *36*, 1−19.
(30) Nelder, J. A.; Mead, R. A simplex method for function minimization. *Comput. J.* **1965**, *7*, 308.
(31) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Daniel, T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(32) Fan, H.; Schneidman-Duhovny, D.; Irwin, J. J.; Dong, G.; Shoichet, B. K.; Sali, A. Statistical Potential for Modeling and Ranking of Protein-Ligand Interactions. *J. Chem. Inf. Model.* **2011**, *51*, 3078−3092.
(33) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48*, 2534−2547.
(34) Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* **2007**, *35*, W522−W525.
(35) DeLano, W. L. *The PyMOL molecular graphics system*; Schrodinger Inc.: Cambridge, MA, 2002.
(36) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035.
(37) Irwin, J. J.; Shoichet, B. K. ZINC-a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(38) Jorgensen, W. L. *QikProp*; Schrödinger LLC: New York, 2006.
(39) Ruppert, J.; Welch, W.; Jain, A. N. Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci.* **1997**, *6*, 524−533.
(40) Baber, J. C.; Thompson, D. C.; Cross, J. B.; Humblet, C. GARD: a generally applicable replacement for RMSD. *J. Chem. Inf. Model.* **2009**, *49*, 1889−1900.
(41) Li, X.; Li, Y.; Cheng, T.; Liu, Z.; Wang, R. Evaluation of the performance of four molecular docking programs on a diverse set of protein-ligand complexes. *J. Comput. Chem.* **2010**, *31*, 2109−2125.
(42) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of several molecular docking programs: pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2009**, *49*, 1455−1474.
(43) Spitzer, R.; Jain, A. N. Surflex-Dock: Docking benchmarks and real-world application. *J. Comput.-Aided Mol. Des.* **2012**, Epub ahead of print.
(44) Novikov, F. N.; Stroylov, V. S.; Zeifman, A. A.; Stroganov, O. V.; Kulkov, V.; Chilov, G. G. Lead Finder docking and virtual screening evaluation with Astex and DUD test sets. *J. Comput.-Aided Mol. Des.* **2012**, Epub ahead of print.
(45) von Korff, M.; Freyss, J.; Sander, T. Comparison of ligand- and structure-based virtual screening on the DUD data set. *J. Chem. Inf. Model.* **2009**, *49*, 209−231.
(46) Alberts, I. L.; Todorov, N. P.; Dean, P. M. Receptor flexibility in de novo ligand design and docking. *J. Med. Chem.* **2005**, *48*, 6585−6596.
(47) Fischer, B.; Merlitz, H.; Wenzel, W. Increasing diversity in in-silico screening with target flexibility. *Comput. Life Sci.* **2005**, 186−197.
(48) Neves, M. A.; Totrov, M.; Abagyan, R. Docking and scoring with ICM: the benchmarking results and strategies for improvement. *J. Comput.-Aided Mol. Des.* **2012**, Epub ahead of print.

(49) Ehlert, F. J.; Delen, F. M.; Yun, S. H.; Liem, H. A. The interaction of amitriptyline, doxepin, imipramine and their N-methyl quaternary ammonium derivatives with subtypes of muscarinic receptors in brain and heart. *J. Pharmacol. Exp. Ther.* **1990**, *253*, 13−19.
(50) Richelson, E.; Nelson, A. Antagonism by antidepressants of neurotransmitter receptors of normal human brain in vitro. *J. Pharmacol. Exp. Ther.* **1984**, *230*, 94−102.
(51) Meyer, J. M.; Ejendal, K. F.; Avramova, L. V.; Garland-Kuntz, E. E.; Giraldo-Calderon, G. I.; Brust, T. F.; Watts, V. J.; Hill, C. A. A "Genome-to-lead" approach for insecticide discovery: pharmacological characterization and screening of *Aedes aegypti* D(1)-like dopamine receptors. *PLoS Negl. Trop. Dis.* **2012**, *6*, e1478.
(52) Gatica, E. A.; Cavasotto, C. N. Ligand and decoy sets for docking to G protein-coupled receptors. *J. Chem. Inf. Model.* **2012**, *52*, 1−6.
(53) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **2008**, *153* (Suppl 1), S7−26.
(54) Tarasov, D.; Tovbin, D. How sophisticated should a scoring function be to ensure successful docking, scoring and virtual screening? *J. Mol. Model.* **2009**, *15*, 329−341.
(55) Carlsson, J.; Coleman, R. G.; Setola, V.; Irwin, J. J.; Fan, H.; Schlessinger, A.; Sali, A.; Roth, B. L.; Shoichet, B. K. Ligand discovery from a dopamine D3 receptor homology model and crystal structure. *Nat. Chem. Biol.* **2011**, *7*, 769−778.
(56) Mysinger, M. M.; Weiss, D. R.; Ziarek, J. J.; Gravel, S.; Doak, A. K.; Karpiak, J.; Heveker, N.; Shoichet, B. K.; Volkman, B. F. Structure-based ligand discovery for the protein-protein interface of chemokine receptor CXCR4. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 5517−5522.

4.5.4.- Cribado Virtual de OXA-24

La enzima bacteriana β-lactamasa de clase D OXA-24 de *Acinetobacter baumannii* (Santillana, Beceiro, Bou, & Romero, 2007) hidroliza antibióticos con anillos β-lactámicos de tipo carbapenem lo que le confiere resistencia a los tratamientos con este tipo de fármacos. Con el fin de evitar su degradación los antibióticos se suelen suministrar en combinación con inhibidores de las β-lactamasas. La actividad carbapenemasa presentada por estas proteínas es de especial relevancia ya que degradan antibióticos destinados a atacar patógenos ya de por sí multi-resistentes El mecanismo de acción de OXA-24 viene dado por eventos de acilación/deacilación de serinas del sitio activo, regulados por aguas catalíticas y un residuo modificado denominado KCX, una lisina carboxilada (Figura 27). Los inhibidores sulfonados cristalizados con OXA-24 se unen covalentemente a la proteína a través de la serina del motivo catalítico Ser-X-X-KCX bloqueando así la capacidad de la bacteria de degradar el antibiótico suministrado. El centro activo posee una carga positiva global y se encuentra en la interfaz de dos subdominios adoptando la apariencia de una cavidad extendida con una barrera hidrofóbica (6 Å de diámetro del túnel) que regula la forma y naturaleza química de los antibióticos que pueden acceder al sitio activo (Figura 27).

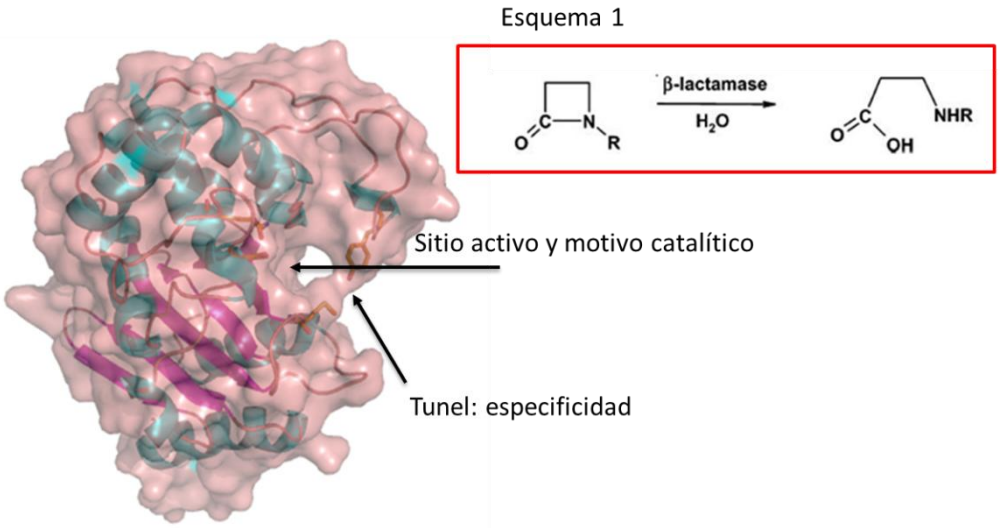


Figura 27. RX de la proteína OXA-24 [PDB ID: 2JC7], localizando el motivo catalítico Ser-X-X-KCX, sitio de unión a anillos β-lactámicos de tipo carbapenem, y evidenciando la especificidad de la proteína por dichos antibióticos adquirida mediante mutaciones en las posiciones TYR112 y MET223 que generaron un túnel que controla el acceso al sitio catalítico. En la caja se ilustra un esquema simple de un anillo β-lactámico y su degradación.

La autora de esta tesis llevó a cabo una búsqueda *in silico* de inhibidores para la β-lactamasa de clase D OXA-24 mediante el VS de quimiotecas usando la plataforma VSDMIP (Cabrera et al., 2011) y la base de datos ZINC (Irwin & Shoichet, 2005). Disponíamos de 5 estructuras de partida: forma apo, forma apo más el residuo modificado KCX, y forma holo

unida a tres inhibidores diferentes. En primer lugar preparé la proteína, parametrizando la lisina carboxilada KCX según el campo de fuerzas de AMBER y realicé el *docking* de inhibidores conocidos para comprobar que los ligandos adquirían una conformación y pose correctas a pesar de que no considerábamos su enlace covalente con la proteína. A continuación, llevé a cabo el cribado virtual de 166 análogos de 3 inhibidores conocidos de OXA-24 para los cuales existe estructura cristalográfica en complejo con dicha proteína. De ellos, 45 complejos fueron refinados a lo largo de 2 a 4 ns de MD. Obtuvimos 35 candidatos o *hits* que se mantenían estables dentro del sitio activo durante la dinámica. En base a la energía de los complejos, a la solubilidad y a la eficacia del ligando (LE), finalmente seleccioné 5 moléculas que fueron propuestas para los análisis *in vitro* e *in vivo.* Los resultados preliminares señalaron actividad inhibidora para la mejor de las moléculas propuestas (*i.e.* la de menor energía en el ranking de las 5), aunque sólo a altas concentraciones (micromolar). Por ello, dicha molécula se encuentra actualmente en fase de optimización.

4.6.- Caracterización de mutaciones puntuales de proteínas involucradas en la resistencia a fármacos de pacientes con VIH e identificación de sus variantes minoritarias

4.6.1.- Introducción y aportación del autor

La terapia personalizada, que incluye el diseño de métodos de evaluación que permitan adaptar cada tratamiento a un paciente específico, es uno de los retos de la medicina moderna. Los *microarrays* de expresión de ADN se vienen aplicando con éxito desde hace varios años en enfermedades como la leucemia u otros tipos de cáncer mediante la comparación de la expresión diferencial de un conjunto de genes entre un tejido sano y uno enfermo (Trevino et al., 2007). Su aplicación a la detección y caracterización de poblaciones víricas sin embargo es algo más reciente (Chiu et al., 2008).

Dado que para ciertas enfermedades víricas disponemos de un conjunto de fármacos alternativos, la caracterización de las mutaciones puntuales de variantes víricas que presentan los pacientes en el marco de la terapia personalizada permitirían seleccionar el fármaco o los fármacos más adecuados para el tratamiento de un determinado paciente, evitando los posibles efectos nocivos de aquellos para los que presente resistencia *a priori*. Asimismo, la detección de variantes víricas minoritarias ayudaría a adelantarnos a posibles resistencias *a posteriori*, ya que cuando atacamos a las variantes mayoritarias con un fármaco las variantes minoritarias pueden adquirir importancia convirtiéndose en dominantes y promoviendo el rebrote de la enfermedad. Por tanto, el conocimiento detallado de una población vírica a través de sus variantes genéticas ayudará en el pronóstico, diagnóstico y tratamiento individualizado de la enfermedad asociada. Debido a que la evolución puede influenciar la patogénesis de los virus y su resistencia a los tratamientos. En el artículo que se presenta a continuación se ha diseñado un método para estudiar la variabilidad y la dinámica genética de una población viral permitiendo discriminar entre posiciones *wildtype* (WT) y mutaciones de resistencia a inhibidores conocidos.

La autora de la tesis ha contribuido al artículo en preparación (Martín V. et al, 2013) con la implementación de un protocolo estadístico para la cuantificación y control de calidad de los datos provenientes de *microarrays* de expresión de oligonucleótidos para identificar mutaciones puntuales e inserciones de la proteasa (PR) y la retrotranscriptasa (RT) en las secuencias del virus de la inmunodeficiencia humana (VIH), además de estimar el nivel de detección de sus variantes minoritarias presentes en muestras de pacientes.

Estos *microarrays* presentan características particulares que hacen que las técnicas habituales de normalización y análisis estadístico diseñadas para los *microarrays* (*microarrays* de AFIMETRIX por ejemplo), no sean adecuadas. Algunas de sus particularidades residen en el escaso número de datos, así como en la longitud de las sondas a hibridar (15 nucleótidos). Para ello, en primer lugar se procede a analizar un conjunto de muestras con variantes o clones puros y mezclas de dos clones conocidos a concentraciones definidas. En base a las señales obtenidas y con el objetivo de reducir principalmente los falsos positivos durante la clasificación final, las señales positivas y negativas se normalizan de modo independiente y se definen filtros a varios nivel: **(1)** pocillo o *spot*, **(2)** sonda y **(3)** *microarray* completo, eliminando gran parte del ruido de las muestras. En segundo lugar evaluamos la capacidad predictiva del método usando un conjunto de muestras de pacientes de las que conocemos experimentalmente (por secuenciación) las secuencias y proporciones relativas de 20 variantes de la población viral de cada paciente.

En el conjunto de entrenamiento de clones puros la capacidad predictiva, entendida como el porcentaje de clasificados correctamente (TP y TN) respecto al total, alcanzó el 96,33% cuando clasificamos los datos tras el filtrado de 4 pasos propuesto en el artículo respecto al 93,04% usando los datos sin filtrar. En el caso de las mezclas de dos clones puros los datos varían de 95,32% a 97,64% en la mezcla 1.95c9-2.94c64 y de 91,15% a 94,26% en la mezcla pWT-pINS entre un protocolo sin filtrado y otro con 4 filtros secuenciales respectivamente. En este conjunto la sensibilidad de detección se estableció entorno al 10% a pesar de la variabilidad mostrada entre las sondas con hibridación diferencial. En el conjunto de evaluación formado por muestras de pacientes multi-resistentes a inhibidores de PR y RT llegamos a clasificar correctamente el 93,53% de los datos de PR (84,42% sin filtros) y el 89,24% de RT (85,27% sin filtros). La clasificación de las variantes minoritarias en muestras de pacientes (un 7,5% de los datos totales) ha sido correcta en el 85,7% de los casos.

La inclusión del filtrado de 4 pasos así como la normalización diferencial de las señales positiva y negativa han mejorado la capacidad predictiva del clasificador en todos los conjuntos de datos presentado manteniendo sus valores en un rango de entre un 90% y un 98% de fiabilidad dependiendo del caso, haciendo de éste un método adecuado en el pronóstico, el tratamiento y el seguimiento en el tiempo de la población viral de un determinado paciente.

Artículo 7

VIROCHIP 4.0: an efficient and fast genotyping platform for the identification of drug-resistance mutations and insertions in HIV populations.

Verónica Martín (CBMSO; actual: INIA), Celia Perales (CBMSO), María Fernández-Algar (CAB), **Helena G. Dos Santos** (CBMSO), Patricia Garrido (X-Pol Biotech), María Pernas (ISCIII, Majadahonda), Víctor Parro (CAB), José Alcamí (ISCIII, Majadahonda), José Luis Torán (CNIC), David Abia (CBMSO), Esteban Domingo (CBMSO), Carlos Briones (CAB).

Abstract

The evolution of viral quasispecies can influence viral pathogenesis and the response to antiviral treatments. Microarray technology, originally developed for highly parallel examination of gene expression is considering as a potential tool in prognosis and diagnosis. DNA microarrays are becoming increasingly useful for the analysis of gene expression and single nucleotide polymorphisms (SNPs).

In this study, using Human Immunodeficiency Virus type 1 (HIV-1), as a model study, we demonstrated the advantages of the oligonuleotide-based microarray approach for the rapid and accurate detection and discrimination between wild type and mutants of protease (PR) and/or retrotranscriptase (RT) from HIV-1 at a specific genome position. A rapid and reliable method for the identification of PR and/or RT from HIV-1 strain B has been developed. The genotype-specific oligonucleotides immobilized on the surface of glass slides were selected to bind to the multiple drug-resistance mutations developed by PR- and RT-HIV in multidrug treatments of infected patients. HIV-1 cDNA was amplified in a PCR with specific primers to PR- and RT-HIV-1. A second round of nested PCR amplification was performed. The use of one primer containing 5'-phosphorilated allowed us to obtain single-stranded DNA susceptible to be labelled by Alexa 647 Fluor. The identification of HIV-1 quasispecies genotype was based on hybridization with several individual genotype-specific oligonucleotides present in the array. This approach combines the high sensitivity of PCR with the selectivity of DNA-DNA hybridization. The specificity of oligonucleotide microchip hybridization was evaluated by testing different PR- and RT-HIV-1 cloned into a plasmid and more than 50 HIV-1 samples extracted from patients for whom genotypes were previously determined by conventional methods. Analysis of the coded specimens showed that this microarray-based method is capable of unambiguous identification of 82 drug resistant mutations and insertions in PR- and RT-HIV.

Introduction

The different variants of human immunodeficiency virus type 1 (HIV-1) present in infected individuals have been described as quasispecies, complex distributions of related, non-identical viral genomes (Coffin, 1995; Eigen and Biebricher, 1988; Holland, de La Torre, and Steinhauer, 1992). In HIV-1, family of retroviruses with single-stranded RNA genome, population complexity underlies evolution of a number of phenotypic traits. This plasticity of phenotype allows the virus to occupy a large adaptive landscape from which novel phenotypes may readily emerge, including rapid selection of mutant viruses with decreased sensitivity to antiretroviral inhibitors in patients treated with reverse transcriptase (RT) or protease (PR) inhibitors (Menéndez-Arias, 2002). Mutations related to drug resistance may either pre-exist in HIV-1 quasispecies or be generated de novo and rapidly selected in infected patients. These variants are generated continuously due to the high replication rate of HIV-1 (Perelson et al., 1996), the high frequency of recombination among viral genomes (Burke, 1997), and the lack of proof-reading activity of the viral reverse transcriptase (RT) (Mansky and Temin, 1995).

The quasispecies dynamics predicts that the simultaneous administration of multiple antiretroviral inhibitors directed to independent viral targets should minimize and delay selection of inhibitor-resistant HIV-1 mutants (Domingo, 1992; Domingo and Holland, 1992; Wodarz and Nowak, 1999). Current HIV therapies inhibit the viral replication process at the binding and entry stage (fusion inhibitors), the reverse transcription stage (nucleoside and no nucleoside reverse transcriptase inhibitors [NRTIs and NNRTIs, respectively]), or the protein cleavage stage (PIs). Inhibitors of coreceptor binding, integration, and maturation are in clinical trials. Benefits of combination therapy, including highly active antiretrovial therapy (HAART) involving three or more inhibitors, have been amply documented in clinical trials (Gulick et al., 1997; Hammer et al., 1997). Yet, either because of a history of monotherapy or other no suppressive treatment regimens, or because HIV-1 replication is not completely inhibited under HAART, viruses including constellations of mutations conferring resistance to multiple inhibitors are increasingly frequent (Schmit et al., 1996; Schmit et al., 1998; Wong et al., 1997). The emergence of HIV-1 mutants resistant to antiviral inhibitors is a complex process which depends on a number of interlinked parameters: mutation rates, number of mutations needed to confer different degrees of resistance, viral replication rounds, viral population size, relative fitness of the different variants, and effective inhibitor concentrations at different body sites where virus replication takes place (Domingo, 1999; Wodarz and Nowak, 1999). Current evidence suggests that a few critical mutations in the PR and RT are sufficient to render HIV-1

partially resistant to most antiviral inhibitors in use or under development (Palmer, Shafer, and Merigan, 1999).

The evolution of viral quasispecies can influence viral pathogenesis and the response to antiviral treatments (Pawlotsky, 2006). The extensive use of anti-retroviral therapy and multiple combination treatments has favoured the emergence of novel patterns of mutations conferring multi-drug resistance. HIV drug resistance is presently a major factor responsible for antiviral drug failure in patients. Resistance testing can help to make intelligent choices for changing highly active antiretroviral therapy. The control of diseases associated with highly variable RNA viruses requires close monitoring of the variant virus types that periodically dominate in viral populations.

Despite the fundamental role this dynamic polymorphism plays in processes having important practical implications, current experimental methods for its analysis do not easily reveal the full picture. First, any protocol that includes cDNA cloning is inadequate unless it involves generation of a large representative subset of independent clones. Classical methods for evaluation of genomic heterogeneity of RNA viruses are generally very labour-intensive and poorly adapted for high-throughput analysis. The traditional approach for the investigation of the mutant spectrum of HIV-quasispecies involves the sequence analysis of a representative number (10 to 30) of molecular clones derived from the amplified viral population (Briones et al., 2000; Kijak et al., 2002; Paolucci et al., 2001; Westby et al., 2006). Nevertheless, this is a very expensive and time-consuming method that would require the analysis of 1,000 molecular clones to characterize the quasispecies at a resolution of 0.1%. Therefore, the clonal analysis is an unrealistic approach for most clinical laboratories. In addition to the most informative but tedious sequencing of a large number of viral clones, there are methods including PCR-restriction fragment length polymorphism (RFLP) (Balanant et al., 1991), mutant analysis by PCR and restriction enzyme cleavage (Chumakov et al., 1991), and real-time quantitative PCR which allows the characterization of minor subpopulations of drug-resistant HIV variants in sero-converters and in pre-treated patients (Allers et al., 2007; Bergroth, Sonnerborg, and Yun, 2005; Charpentier et al., 2004; Hance et al., 2001; Izopet et al., 2002; Lecossier et al., 2005; Metzner et al., 2003; Metzner et al., 2005). This approach has led to the detection of minority variants representing from 0.05% to 1% of the total amplified quasispecies. During the last decade novel methodologies, based on rapid and highly sensitive assays, have been developed and are currently permeating the community of HIV-1 clinical research. Different variants of the heteroduplex mobility and heteroduplex tracking assays have been used to detect minority subpopulations comprising 1-5% of the amplified viral

quasispecies (Barlow, Green, and Clewley, 2000; Doukhan and Delwart, 2001; Kapoor et al., 2004; Resch et al., 2001).

Also, DNA microarray technologies have been applied to high-throughput viral genotyping (Ivshina et al., 2004; Martin et al., 2006; Wang et al., 2002) and can be adapted or combined with mass spectrometry to simultaneously inform one about different resistance mutations present in genomes represented in the range of 1% of the total amplified quasispecies (Amexis et al., 2001; Cherkasova et al., 2003; Gerry et al., 1999). Recently an alternative sequence-based method, involving parallel allele-specific sequencing, has been developed to allow the detection of minority HIV-1 drug-resistant variants present at levels ranging from 0.01% to 0.1% of the amplified sample (Cai et al., 2007). Areas of research requiring methods devoid of these short-comings include screening for emerging viruses, monitoring of viral evolution indifferent natural hosts, and safety of live viral vaccines.

Hybridization with microarrays of immobilized DNA or oligonucleotides that contain thousands of individual probes appears to be suitable for parallel analysis of a large number of viruses, bacteria, and bacterial factors (Chizhikov et al., 2001; Chizhikov et al., 2002; Lapa et al., 2002; Volokhov et al., 2002; Wang et al., 2002). The methods were based on either short highly specific oligonucleotide probes (oligoprobes) or longer degenerate probes that covered a wide range of viral diversity.

We describe here the validation of a DNA microarray by using RT and PR-HIV-1 cloned into different plasmids as well as clinical HIV-1 samples isolated from infected patients.

Materials and Methods

Experimental datasets

The training set includes microarray hybridization data from a pool of pure HIV-1 clonal samples (268 hybridization experiments) and mixtures of 2 pure clones (110 hybridizations) at different ratios: 0/100, 1/99,5/95, 10/90, 50/50, 90/10, 95/5, 99/1 and 100/0 (in %). The test set was composed by hybridization data of HIV-1 quasispecies extracted from 50 clinical samples of multidrug-resistant patients (89 hybridizations).

Microarray design and printing

One hundert fifty-eight DNA oligonucleotides, corresponding to the HIV genomic region encoding residues 2255 to 3872 were designed and synthesized (Sigma). They included

a ´C6 amino linker´ [NH_2 $(CH_2)_6$] at their 5'-end, followed by an oligo $(dT)_{15}$ spacer or $(dTCC)_5$ and the specific 15-mer sequence; the oligonucleotides were purified by HPLC. The oligonucleotides (Figure 3) were selected to have a similar melting temperature when annealed to a complementary sequence, and included the queried nucleotide at the central region of the specific 15-mer. Two conserved HIV sequences, [5'-C6-T_{15}GATACAGGAGCAGAT-3' (ICH-PR3) and 5'-C6-T_{15}GTTGCACTTTAAATTTT-3' (ICH-PR4)], within the PR-coding region [5'-C6-T_{15}ATGGCCATTGACAGA-3' and (ICH-RT3) 5'-C6-T_{15}ATCTTAGAGCCTTTTA-3' (ICH-RT4)] within the RT-coding region) were used as positive controls for the hybridization (ICH, Internal Control HIV). Two unrelated oligonucleotides (5'-C6-T_{15}-CAAATCCCCGCGTGC-3' and 5'-C6-T15-CAAATCCTCGCGTGC-3', termed G142-15r (corresponding to the complementary sequence of the FMDV VP1 coding region at genomic residues 3625 to 3639) and E142-15r (corresponding to the same region with the nucleotide 3632 mutated from C to T) were used as negative controls. Spots with spotting solution with no nucleotide (H_2O-ss in Figure 4A) were used as negative controls. The oligonucleotides were diluted in 1x spotting solution (Telechem-Arrayit) at a 50 μM final concentration, and spotted onto super-epoxy-coated glass slides (Telechem-Arrayit).

Microarrays containing 332 spots were printed by means of a GMS 417 DNA arrayer (Affymetrix) defining four grids per slide. Each oligonucleotide was spotted in duplicate dots 150 μm in diameter, with a center-to-center distance of 250 μm (Figure 4A).

Preparation of targets DNAs

From plasmids

HIV sequences were amplified using Expand High Fidelity polymerase (Roche), as specified by the manufacturers. PCR amplifications of the PR-coding region was carried out using the sense primer 5' PROT HindIII (5'-TCAGAGCAGACCAGAGCCAACAGCCCCACC -3'); corresponding to HIV genomic residues 2138 to 2167 (phosphorylated at its 5'-end) and antisense primer 140RD (5'-CCAGGGATTAGATATCAGTACAATG-3'), complementary to positions 2969 to 2993 of the HIV genome. PCR amplifications of the RT-coding region were carried out using the sense primer 55F (5'-CAAAAATTGGGCCTGAAAATCC-3'); corresponding to HIV genomic residues 2694 to 2715 or RT1 (5'-CCAAAAGTTAAACAATGGCCATTG-3') corresponding to HIV genomic residues 2606 to 2629, phosphorylated at their, 5'-end, respectively; and antisense primers 153RD (5'-GGAAAGGATCACCAGCAATA-3'), 12RD (5'-AACATCAGAAAGAACCTCCA-3'), 13RD (5'-CCTTTGGATGGGTTATGAAC-3') or 20RD (5'-

GCCAGAAAAGGACAGCTGGACTGT-3'), complementary to HIV genomic residues 3009 to 3028, 3210 to 3229, 3232 to 3251 and 3289 to 3312, respectively. (Scheme of primers used in Figure 1).

From infected patients

HIV-1 DNA was kindly provided by different hospitals (see Origin of HIV samples). External PCR amplification of the cDNA was performed using Expand High Fidelity polymerase (Roche), as specified by the manufacturers. It was carried out using the sense primer 5'PROT1 (5'-AGGCTAATTTTTTAGGGAAAATCTGGCCTTCC-3'), corresponding to HIV genomic residues 2080 to 2111 and antisense primer RT3333R (5'-CAATGACATACAGAAGTTAGTGGG-3'), complementary to positions 3313 to 3336 of the HIV genome.

Due to the small amount of DNA generated by reverse transcription and PCR amplification (RT-PCR), a nested PCR was required. Nested PCR amplifications of the PR-coding region were performed with the primers 5' PROT HindIII (5'-TCAGAGCAGACCAGAGCCAACAGCCCCACC -3'); sense primer, corresponding to HIV genomic residues 2138 to 2167 (phosphorylated at its 5'-end) and 140RD (5'-CCAGGGATTAGATATCAGTACAATG-3'); antisense primer, complementary to positions 2969 to 2993 of the HIV genome. Nested PCR amplifications of the RT-coding region were performed with the sense primer RT1 (5'-CCAAAAGTTAAACAATGGCCATTG-3'), corresponding to HIV genomic residues 2606 to 2629 (phosphorylated at its 5'-end) and antisense primers 153RD (5'-GGAAAGGATCACCAGCAATA-3'), 12RD (5'-AACATCAGAAAGAACCTCCA-3'), 13RD (5'-CCTTTGGATGGGTTATGAAC-3') or 20RD (5'-GCCAGAAAAGGACAGCTGGACTGT-3'), complementary to positions 3009 to 3028, 3210 to 3229, 3232 to 3251 and 3289 to 3312, respectively (Scheme of primers used in Figure 1). In addition to reactions with undiluted DNA, reactions were also performed from diluted samples 1/10 and 1/100, being positive. Some of the PCR products were analyzed by nucleotide sequencing (consensus sequence) using the Big Dye Terminator Cycle Sequencing Kit (Abi Prism, Perkin Elmer) and the automated sequencers ABI 373 or ABI 3700, to ensure the presence of the corresponding mutations. The phosphorylated strand was specifically degraded using lambda exonuclease (New England Biolabs), and the resulting single-stranded DNA was labeled with Alexa Fluor 647 using the U-21660 Ulysis Nucleic Acid Labeling Kit (Molecular Probes). The labeled DNA was used as target in the hybridization with the probe oligonucleotides on the microarrays (Figure 2).

Molecular cloning

The composition of the patient quasispecies was determined by clonal analysis. For molecular cloning, the products of the nested PCR-amplified DNA the "TA cloning kit for sequencing" (Invitrogen) was employed. The sequences of 20 molecular clones of each sample were analyzed using automated sequencing provided by Secugen company. 2568 clones were analyzed. The analysis of these sequences was performed with the Lasergene and BioEdit informatic programs.

Hybridization and scanning

Immediately before hybridization slides washed for 2 min. at room temperature with 2x sodium saline citrate (SSC), 0.1% lauroylsarcosine, and for an additional 2 min. with 2xSSC at room temperature, to remove unbound DNA and components of the printing buffer. The oligonucleotides were denatured by placing the slides for 2 min. in distilled water at 100ºC and cooled for 10 sec. at room temperature. Then the oligonucleotides were fixed by plunging the slides into ice-cold 100% ethanol for 2 min. Finally the slides were centrifuged for 1 min. at 500 xg (Minicentrifuge Arrayit). Microarrays were incubated in a hybridization chamber (Genetix) with 20 μl of hybridization buffer (6xSSC, 0.5% SDS, 1% BSA) under a 24 × 24 mm cover slip, and bathed at 42ºC for 45 min. Then the microarrays were washed with distilled water, and dried by a brief centrifugation.

The hybridization with the labeled DNA was carried out in hybridization buffer (50ºC) and with the required amount of target (0.3 pmoles Alexa Fluor 647 equivalent to 50 ng). After 3 hours incubation in the hybridization chamber, the slides were washed for 5 min. in 2×SSC, 0.1% lauroylsarcosine, followed by 5 min. in 2×SSC, and finally rinsed 10 sec. in 0.2×SSC, and 5 min. in distilled water, at 45ºC. The slides were dried by spinning for 1 min. at 500 xg and, finally, scanned using a G2565AA/G2565AB Scanner (Agilent). The Agilent and Scan Array Express (Perkin Elmer Life Sciences) analysis software was used for reading and quantifying the hybridization images. The reproducibility of the method was assessed by comparing the results of at least five different hybridization experiments for each PR or RT cloned into a plasmid.

Microarray quantification and quality control of hybridization signals program development

Image analysis and quantification

Experimental quantification of microarray signals was carried out using two different high resolution microarray scanners: GenePix 4000B (Molecular Devices) and ScanArray

Express (Perkin Elmer). Images were processed with their own software (GenePix Pro 6.0.0.68 and ScanArray Express 2.0, respectively) in order to calculate, among other variables, the median intensity for each spot and its respective local background.

Quality control of arrays and spots

A quality control protocol has been implemented to exclude low quality hybridization signals for further analyses, thus making microarray data analysis more robust. Quality is evaluated at three different stages during the data processing protocol: complete microarrays, single spots and specific probes. Probe oligonucleotides spotted onto the microarrays explore a variable number of codons for each position of interest of the PR and RT regions of the pol gene. On average, there is a probe for each WT codon sequence and two probes for alternative mutant versions of it. Thus, from a theoretical point of view, the expected number of positive signals for pure clonal samples should be 1/3 of the probes included in each microarray. This criterion allows us to discard the arrays that present values higher than 12 and 35 for PR and RT respectively. Above this cut-off each microarray accumulates more than 5 and 20 errors for PR and RT respectively. Next, in order to remove the noise arising from hybridization experiments which is not due to biological variation, spot filters based on background intensity and spot diameter measures have been included before the normalization step. The election of the filters was based on a preliminary study of the distribution of quantified data binned as correctly classified or not. Spots which did not satisfy this quality control were not considered for further analysis.

Normalization

Normalization of the data has been recognized as a necessary pre-processing step in a variety of high-throughput technologies, including DNA microarrays. All known normalization methods rely on assumptions about data features that are expected to be invariant across samples. Different normalization methods have been developed for gene expression microarrays (Bolstad et al., 2003), but there is no consensus on their relative performance on genotyping microarrays containing a small number of oligonucleotide probes. Nevertheless, the incorporation of an accurate normalization method in the analysis is required to trace the array-to-array variability.

The selected raw fluorescence value of each spot was the median foreground signal after subtraction of its local background. Spots from each target-probe hybridization signal (4 spots per experiment) were clustered together. We assume that some of the probes spotted

onto the microarray will specifically hybridize to their fully complementary sequences present in the (either majority or minority subpopulation of) labelled target DNA, rendering a positive signal. The rest, although presenting a residual hybridization signal, will be called negative signal or noise. Therefore, normalization was carried out independently using the following method: first, since two different classes of signals (positive or negative) are expected, they were clustered into 2 groups using the K-means algorithm (Forgy, 1965). After removing the outliers from the positive group by applying the Grubb's test (Grubbs, 1950) with a significance level of 0.05, the mean of the positive signals was obtained. Since most of the hybridization experiments have been performed using labelled DNA molecules corresponding to either the PR or the RT-coding sequences of the *pol* gene, two microarray regions (containing the PR and RT probes, respectively) were normalized separately by using their own mean positive signal. These values were used as the normalization factor of each microarray region by applying the following function:

$$\|PS\| = \sqrt{\frac{\sum_{i=1}^{N^{RP}} I_i^{RP}/N^{RP}}{\sum_{i=1}^{N^{PS}} I_i^{PS}/N^{PS}}}$$

Where $\|PS\|$ is the normalized probe signal, $N^{RP}$is the total number of replicas per probe (typically, 4), I^{RP} is the intensity of a replica probe, N^{PS} is the total number of positive probe signals per array region (either PR or RT) and I^{PS} is the intensity of a positive probe signal within such a region.

Classification performance

In order to evaluate the classification accuracy of the method, tables showing the theoretical hybridization signal of each probe with each hybridized target molecule (pure clonal samples) are required. The so-called 'theoretical hybridization tables' have been generated based on the sequence complementarity between the spotted probes and the (previously sequenced) targets belonging to the training set: a positive hybridization signal is expected when probe and target are fully complementary, while the presence of any single mismatch between them (including those at the 5' or 3' ends of the hybridizing sequences) is assigned to a negative signal. Therefore, partial hybridizations are not allowed in the theoretical tables, and it is assumed that this criterion will assign any incomplete hybridization that resist the washing step (e.g., those including 1 or 2 mutations between probe and target) to false positive signals.

When the hybridized samples are a synthetic mixture of two HIV-1 clones, a 'consensus table' has to be generated based on the percentage of each clone present and a pre-defined detection threshold. For setting such a threshold, we used a collection of mixtures of two clonal samples with known mixture ratios. Once we determine the detection threshold in mixtures, we use this information to generate the theoretical consensus table of clinical HIV-1 samples in which a number of clones (between 20 and 30) have been previously sequenced and aligned. The comparison of this theoretical consensus table with the hybridization signal obtained when hybridizing clinical samples will inform us about the sensitivity of our microarrays for detecting minority subpopulations within the evolving HIV-1 quasispecies.

Probes calibration

Although all the spotted probes have been designed and tested to show similar hybridization temperatures (between 49ºC and 54ºC), not all of them behave equally in hybridization experiments. Thus, in order to classify the observed hybridization intensities, it is necessary to characterize the response profile of each probe when it hybridizes either with its specific target sequence or with an unspecific one.

Calibration data have been collected from the normalized hybridization signals produced by the training set of pure clonal samples. Spot intensities were reorganized into groups of positive or negative hybridization signals for each probe based on the previously calculated theoretical hybridization tables. We assume that both signal and noise data can be modelled by function distributions. In particular, Normal and log-Normal distributions were independently adjusted to positive and negative normalized datasets, respectively**.** Normalized signal values from positive probes are expected to be centred in 1, while negative intensities will have values close to 0, allowing an easy discrimination between positive and negative hybridizations. Distribution parameters obtained from fitted curves per probe will be used in further steps to assess the probability of any observed intensity to belong to a positive or negative hybridization event. For some probes, the characterization of one type of signal (either positive or negative) is missing due to the absence of information in the current training set. In those cases, we use the average information derived from the whole training set of pure clonal samples. To obtain the so-called 'global reference curves' we cluster the available normalized data from the training set into 2 groups (positive and negative signals) based on the theoretical hybridization tables. As for individual curves, positive and negative global signals were adjusted to Normal and log-Normal distributions, respectively.

Quality control of the probes

During the development of consecutive versions of this HIV-1 genotyping microarray, similar probes complementary to certain nucleotide regions of the viral genome (e.g., including the same resistance-associated codon in the context of alternative flanking sequences) have been designed and tested. Because of their good performance, some of the probes containing the same interrogating codon have been maintained in the final version of the microarray (the so-called 'Virochip 4.0'). In order to avoid redundant information, in each hybridization experiment we select the equivalent probe that presents lower overlap between its positive and negative adjusted distribution curves. In other cases, when positive and negative distributions for one probe are highly overlapped it is not possible to correctly classify the observed intensities into positive or negative signals. In order to discard such non discriminant probes, we calculate the overlapping area between both distribution functions for each probe by integral calculation. Probes whose overlap between positive and negative distributions is over 0.25 are discarded.

Classification and evaluation

Once a clinical sample from the test set is hybridized to the microarray, we compute the cumulative probability difference to belong to the positive or the negative distribution for each normalized signal. Probabilities are based on the adjusted curves previously obtained for each probe using the training set. As stated above, if positive or negative parameters for a given probe are lacking, we consider those of the respective global reference curve. When the absolute accumulative probability difference is smaller than 0.05, the signal will be classified as 'undefined' because the chance of wrong classification is high. To evaluate the classification accuracy of the method, once we classify a signal as positive or negative, we check if this classification is in agreement with its expected signal derived from the theoretical hybridization table, thus obtaining the final classification of the signal as true positive (TP), true negative (TN), false positive (FP), false negative (FN) or undefined (UD).

Results

Specificity and sensitivity optimization of HIV microarray

In a first approach, oligonucleotides were designed for the set up of an HIV-1 microarray. They represent RNA sequences encoding RT of HIV-1 strain B. Different variants of the insertion of two different amino acids between positions 69 and 70, named insertions 69 were initially tested. A microarray was printed to analyze the influence of long (15-mer) versus short

(11-mer) oligonucleotides, the presence or absence of $(dT)_{15}$ spacers, and the oligonucleotide concentration. At least 5 or 7 different versions of the microarray were developed and tested, adding and checking more oligonucleotides considered relevant to detect drug-resistance mutations in the PR and RT-HIV-1. A number of conclusions were drawn from the results (not shown).

First, the hybridization signals were weaker with oligonucleotides of 11 residues than with oligonucleotides of 15 residues. The second observation was that oligonucleotides linked through a $(dT)_{15}$ track hybridized more efficiently than those without the track in agreement with previous results (Guo et al., 1994). Figure 3 shows the final selected sequences for the oligonucleotides printed in the array. As seen in this. Figure 3, in some oligonucleotides the $(dT)_{15}$ track was substituted by a $(TCC)_5$ track. In those cases a high number of adenines in the oligonucleotide sequence occasioned a distortional hybridization. Third, the experiments indicated that the amount of oligonucleotide attached at concentrations between 5 and 50 µM was not limiting for detection of fluorescent DNAs. We chose the highest concentration tested for the standard protocol. A scheme of the final position of the oligonucleotides printed is depicted in Figure 4A. Preliminary experiments showed also that hybridization solutions including 50% formamide resulted in poor sensitivity, and that the Unyhib solution (Arrayit) produced results comparable to those obtained with the hybridization solution described in Materials and Methods. To generate labelled targets, two different systems were used: direct labelling with Cy3-dUTP and Cy5-dUTP, and indirect labelling with Alexa Fluor 647; the latter proved easier, more reproducible, efficient and yielded targets showing higher stability.

A step-wise increase of hybridization temperatures, between 45ºC and 58ºC, was tested. Low temperatures resulted in poor microarray performance due to high number of false positives. The optimal point mutation discrimination was obtained between 50ºC and 52ºC. Higher temperatures resulted in a progressive and significant loss of signal. Similar comparisons revealed 45ºC as the most adequate temperature for washing the hybridized microarrays. A scheme of the entire procedure with indication of the steps for which variables were screened is depicted in Figure 1. A design of the positions of the oligonucleotides used to amplify the different samples of the PR and RT genes of HIV-1 and their sequences are shown in Figure 2.

Screening of point mutations of the genomic region encoding PR and RT of HIV cloned into a plasmid

Evaluation of HIV microarray using reference viruses

A total of 42 positions within PR and RT of HIV-1 strain B were analyzed by constructing 15-mer oligonucleotides with the queried nucleotide (drug-resistance mutations) located at position 7 to 11 in each 15-mer (Figure 3). 176 oligonucleotides were spotted in duplicate, distributed in 15 rows and 12 columns per grid (Figure 4A). Two conserved PR-HIV-1 (ICH-PR3 and 4) and another two RT-HIV-1 sequences (ICH-RT3 and 4) were used as positive control for the hybridization. Two unrelated FMDV oligonucleotides (G142-15r and E142-15r) and spots with no nucleotide (H_2O+ss) were used as negative control. The same pattern containing spots with 15-mers corresponding to the different queried and controls mutants, and positive (ICH) and negative controls (FMDV, H_2O+ss) were printed four times per slide.

PCR products obtained with DNA from PRs and RTs-HIV-1 cloned into different plasmids as template and one of the pair of oligonucleotides (see Materials and Methods) show in Figure 1as primers, were treated with lambda exonuclease, and labelled with Alexa Fluor 647 as detailed in Materials and Methods. The labelled single stranded DNA was hybridized in the microarray, as described in Materials and Methods.

An example of hybridization of a wild type PR and RT samples after amplification, labelling and scanning, as explained in Materials and Methods, is shown in Figure 4B. Very good signal intensity was obtained at all wild type positions expected to be positives (heavy yellow background in Figure 4A). The sample hybridized presents a different genotype to the one impressed in the array at position 36-PR as wild type, being no mutant neither, and due to the primers used in the amplifications 236-RT and 238-RT amino acid positions are not included in the target, consequently no positive signal were detected at these wild type positions. In other targets used in this work these positions have the wild type genotype printed in the array or the amplification includes all the positions tested in the chip and it will be perfectly detected in the hybridization with the array.

These results (Figure 4B) indicate a good discrimination between positive and negative signals as well as strong signals in the ICH probes (ICH-PR3 and ICH-RT4) and no signal in any of the negative controls (FMDV-G142-15r, FMDV-E142-15r and H_2O+ss), as expected from the perfect match and mismatch hybridization signals, respectively. A good detection of wild type positions is obtained when these genotypes are a hundred per cent represented in the

quasispecies used for hybridization. No cross hybridization with the respective mutants positions is detected.

Furthermore, to test the mutant positions different targets from PR and RT samples cloned into different plasmids with some mutant positions in the genotype, were hybridized. Figure 5 summarizes these results, each panel represents a different position of the PRs (A) or RTs (B) targets obtained from diverse microarray images, given by the Alexa 647 fluorescence signal, after hybridization, washing and scanning, as detailed in Materials and Methods. Part A of the picture shows the perfect discrimination of eight different mutants of the PR-HIV-1, I46, L46, V48, V54, V71, T82, V84, and M90 (written in red colour in the Figure 5). Each mutant could be identified due to a high signal in the perfect match probe and no signal detected (mismatch) in the rest of the oligonucleotides, wild type or no, printed in the array corresponding to a specific position (written in black in the Figure 5). Twelve different mutants of the RT-HIV-1 were detected in Figure 5B, V62, I75, T75, Ins69a, Ins69b, Ins69c, Ins69d, I108, I184, K211, Y215 and C215.

A perfect detection and discrimination of the checked mutant positions is obtained when these genotypes are a hundred per cent represented in the quasispecies used for hybridization. No cross hybridization with the respective wild type positions is detected. Therefore, an array classification method was developed to evaluate the hybridization signals.

Screening of point mutations of the genomic region encoding PR and RT from HIV infected patients

Evaluation of HIV microarray using PR and RT from HIV infected patients

Each panel represents different positions found mutated in diverse microarray images, given by the Alexa 647 fluorescence signal, after hybridization of different PRs (A) and RTs (B) targets amplified from HIV infected patients, washing and scanning, as detailed in Materials and Methods. In red colour was written the genotype target detected, position showing positive signal (perfect match), and in black the rest of oligonucleotides, wild type or no, printed in the array corresponding to this position, negative signal.

Figure 6A shows the unequivocal detection of twelve mutants in the PR gene from HIV-1 obtained from infected patients, N30, I36, I46, L46, V46, V48, V54, V71, A82, I82, S82, M90. The identification of each mutant correlates with a high signal in the perfect match (written in red) probe and no signal detected in the rest of oligonucleotides, wild type or no, printed in the array corresponding to a specific position (written in black). These are some of the PR

mutants detected in the quasispecies hybridized and showed in Figure 6A, that have not been represented by a hundred per cent in their quasispecies. In those cases it could be see in the hybridized panel others positive positions. The quasispecies showing I46 as positive has a 67% of this mutant in the molecular cloning (21 clones analysed) and a 23% of the wild type genotype (M46b, M46c). The position wild type is represented by two oligonucleotides M46b and M46c (comparing sequence in Figure 3). The wild type genotype variants of this sample hybridize preferably with M46b and poorly with M46c, probably depending on the hybridization kinetics of the two different probes and the target.

I82 mutant variants are present in a 38% in this quasispecies (with a 62% of the wild type genotype [V82b]). Very weak signal is detected in T82 position, expected to be negative. This oligonucleótido has the same mutation (G to A) in the central nucleotide sequence printed in the array (Figure 3) as I82, plus another one in the contiguous nucleotide. We can obtain perfect match hybridization with the eight 5' first nucleotides.

It is shown the perfect and undoubtedly detection of twenty two different mutants of the RTs proceeding from HIV-1 extracted from infected patients, V62, S68, S69a, D69, Ins69a, Ins69j, R70a, I100-2, E101, I108, K211, M151, M178; C181, I184, H188, A190, E190, Y215, Q219, E219 and T238. In those cases we no detect different variants of RTs per quasispecies hybridized.

Comparison between sequencing data and microarray hybridizations of PR and RT from HIV infected patients

The results presented in first column of Table 1 show different positions found mutated in diverse microarray images, given by the Alexa 647 fluorescence signal, after hybridization of different PR (A) and RT (B) targets amplified from HIV infected patients, washing and scanning, as detailed in Materials and Methods. The results of the hybridized quasispecies shown have been formed by different proportion of genotypic variants. All the data extracted from the samples analysed by hybridization in our array are compared with the molecular cloning and with the genotype obtained from de hospital in Table 1

Microarray quantification and quality control of hybridization signals

The theoretical hybridization table that summarizes the matching between probes and target sequences from the training set is shown in Supplementary Fig. S1.

Spots quality control

The first filtering criterion is based on different filters depending on the scanner type used to process the hybridization images. The parameters of the microarrays scanned using the ScanArray platform include the diameter of the spot (DS) and the percentage of foreground signal higher than the background minus its standard deviation (BSD), while microarrays scanned in the GenePix system inform about the number of background pixels (BP). Cut-offs were selected in order to discard most of the FP and FN spots identified in a preliminary classification that included only probes with positive or negative signal upon hybridization with targets from the pure training set. Spots of microarrays scanned with the ScanArray platform are accepted either if DS ≤ 150, or if DS > 150 being BSD ≤ 40: these criteria eliminate FP signals, avoiding the filtering of TN ones (Supplementary Fig. S2). In turn, spots of microarrays analysed with GenePix platform pass the filter if BP > 300, a criterion that discards most of the FP but also some of the TP signals. Using these filters, 0.2 % of all data derived from the training set are discarded, while hybridization data from the test set are discarded at percentages of 1.27% (RT region) and 4.05% (PR region).

Normalization

After normalization by the mean positive signal of a given microarray region (PR or RT) the value ‖PS‖ allowed the separation of positive and negative signals in most of the cases. Positive signals are centred in value 1 of normalized intensities, while negative ones remain close to zero. (Supplementary Fig. S3)

Probes calibration by function adjustment

Parameters of the fitted distribution curves have been obtained for each probe from the normalized signals combined with their expected behaviour derived from the theoretical hybridization table. Global reference curves, together with two examples of calibrated curves, are shown in Figure 7. Global reference curves have been constructed for circumventing the limitation that some of the probes have not been tested with both target versions (those producing positive or negative signals). Indeed, the training set includes 52 probes with both positive and negative curves defined (32.90%), 9 probes (mainly, WT ones) with only positive signals (5.70%) and 97 probes (mainly, mutants) with only negative signals (61.40%). Thus, by generating global curves of positive and negative signals, and combining them with the individual ones when required, we were able to classify signals not previously characterized in the training set. Combination of such global and individual distribution curves was possible

because their parameters are similar: the mean of positive distributions for global and individual fits was 0.922 and 0.908 (with mean standard deviations of 0.328 and 0.298), respectively, while the mean of negative distributions was -3.321 and -3.236 (mean standard deviations of 1.133 and 0.991), respectively.

Probes Filtering

Positive and negative signals presented certain degree of overlapping in some of the probes spotted in the last version of the microarray (Virochip 4.0). We observe that 76% of probes discriminate correctly (overlapping < 10% between the log-normal distribution of the negative signal and the normal distribution of the positive signal), while 24% of probes have overlapped data to a certain extent (>10%). These behaviours are exemplified by probes Y188a and M230-3 in Figure 8.

When a maximum overlap of 25% between curves is allowed, 14 probes (9.09% of the total) are discarded (e.g., M230-3). When two probes have been designed for detecting the same codon in different sequence backgrounds, the probe with the highest overlapping is discarded. This criterion made us to discard an additional 10.13% of the probes of Virochip 4.0. Therefore, the amount of probes selected at the end of this quality control was reduced from 154 (37 of them belonging to the PR region and 117 to the RT) to 124 (29 in PR and 95 in RT) ().

Accuracy in viral genotyping

The overview of the classification of signals before filtering (training and test sets) is shown in Table 2. The main source of errors in both sets (5.09%, and 9.06 to 9.80%, respectively) are FP signals appearing at probe positions not expected to show a signal because the sequences of probe and target are not fully complementary (see Supplementary Fig. S1).In turn, FN signals are scarce in the training set (1.33%), whereas their value increases to 2.72% (RT region) and 5.41% (PR) in the test set. Finally, UD spots are found at levels below 2.15% in both sets.

The classification performance after the 4-step, sequential filtering protocol applied (see Methods) is shown in Table 4. The amount of correctly classified signals belonging to the training set increases to 96.33% after the stepwise filtering. In turn, the test set results correctly classified in the 93.53% and 89.24% of cases when the PR and RT regions were hybridized, respectively. Therefore, the filtering protocol is especially effective for the PR signals of the test set. Among the different kinds of errors, both FP and FN signals are clearly reduced during the stepwise filtering, while UD signals are less affected by the process. The

classification accuracy in the test set after the filtering protocol is shown in Supplementary Fig. S5, where probes still accumulating most of the FP signals (*i.e.*, I46-PR, K65-2, D67a-2 and E219) and FN ones (Y181) are clearly identified.

Regarding the genotyping of clinical samples, we analysed the concentration of errors by probes and test samples (Supplementary Fig. S6). It is evident that most of the errors are accumulated in a limited number of probes, probably due to the fact that some overlapping curves (those with overlapping area <25%) have not been discarded during the stepwise filtering process. This is exemplified by probes R103-3 (overlap 11.3% and 18 errors). Nevertheless, in probes Y188a and V108-2 the source of errors (18 and 20, respectively) seems independent of the overlapping area between fitted curves. In parallel, a subset of the hybridized samples are responsible for most of the errors (e.g., A23, A26, A44 and A55, with 22 errors each) due to the high number of mutations accumulated at positions adjacent to the queried codons in the circulated viruses of these pre-treated patients.

Sensitivity of detection

To allow a preliminary exploration of the microarray sensitivity for the detection of minority subpopulations in HIV-1 quasispecies, we used mixtures of two pure clonal samples at previously known ratios. The mixtures used involved samples 1.95c9/1.94c64 (62 microarrays) as well as pWT/pINS (48 microarrays). The complete theoretical hybridization tables of these pure samples and their mixtures are shown in Supplementary Fig. S7, while the 7 probes at which a differential hybridization is expected between the samples in each mixture is depicted in Figure 9A). Genome ratios in the mixtures were 0/100, 1/99, 5/95, 10/90, 50/50, 90/10, 95/5, 99/1 and 100/0 (in %). The hybridization of such mixtures to the microarray, after the corresponding stepwise filtering process of the signals (whose results are summarized in Supplementary Table S1), produced the results summarized in Figure 9B). Each probe presents a characteristic sensitivity or detection for minority genomes. In the mixture 1.95c9/2.94c64, probes K211-2 and R211-2-WT showed positive signals in all hybridization experiments when their specific target was present at percentages of at least 10%. Furthermore, in 50% of the experiments (31 of 62 microarrays) these probes showed a neat hybridization signal when their specific targets were present at proportions of 5% in the mixture. In turn, using the mixture pWT/pINS, 4 out of the 5 discriminating probes showed a positive signal when their specific target was present at percentages higher than 10%, being D67a-2 the only probe able to detect minority genomes at proportions of 10%. Unfortunately, the probe Ins69c produced FP signals even when it specific target (pINS) was not present in the mixture. Therefore, taking

into account the limitation in the number of pure clonal samples available for this section of the work, as well as their differential detection thresholds, we can set an experimental cut-off of 10% as the preliminary sensitivity of the microarrays. This value was used to build the theoretical hybridization tables of clinical samples whose quasispecies composition can be previously analysed by clonal sequencing. The results obtained with the mixture 1.95c9/2.94c64 suggest that the maximum sensitivity of the microarray for detecting minority genomes in binary mixtures can reach a value of 5%.

Detection of minority variants in clinical HIV-1 samples

Before performing the hybridization experiments, the percentage of genetic variants present in the quasispecies at each queried codon was quantified by analysing the sequences of 13 to 34 clones isolated from each of the 50 selected clinical samples. After sequence alignment, the percentage of clones containing each codon was quantified, and the results were graphically depicted for the PR (Figure 10A) and RT (Figure 10B) regions. Of the 970 events of perfect probe-target matching in whole test set, most of them corresponded to signals produced by clonal sequences accounting for 90 to 100% of the quasispecies (Figure 10C). Nevertheless, 73 of the probe-target complementarities (7.5%) corresponded to minority sequences present at proportions lower than 20% in their corresponding sample. The distribution of these minority subpopulations in three intervals (1.00–4.99%, 5.00–9.99% and 10.00–19.99% of the quasispecies) is shown in Table 5. The theoretical hybridization tables corresponding to these clinical samples were constructed, considering that the sensitivity of detection of the microarray was previously set to a value of 10% (Supplementary Fig. S8).

In parallel, all the clinical samples were hybridized to the microarray, and the signals obtained were filtered using the stepwise protocol previously described. This allowed the comparison between the theoretical and the experimental hybridization data, thus showing the classification accuracy of the microarray (Supplementary Fig. S9). Overall, 72.1% of all the expected positive signals in PR probes, and 79.5% in RT ones were detected by the microarray (Figure 4). In the PR region, minority subpopulations in the interval 1.00–4.99% could not been detected, while those represented at proportions of 10–49.99% where correctly detected in 85.7% of the cases. In turn, probes corresponding to the RT region showed higher sensitivity for detecting minority subpopulations at proportions of 1.00–4.99% (41.2% of the cases), 5.00–9.99% (53.8%), and 10.00–49.99% (79.2%). Only 5 of the probes failed in the detection of variants present at proportions ≥ 50% (L46-PR, V71-PR-2, V62, N68 and I106-2), and 2 FN signals were produced in probes I108-2 and T215b-2 when hybridizing with targets that

contained their complementary sequences at proportion ≥ 99%. Noticeably, no FP signals were detected in any of the hybridization experiments involving clinical samples.

Regarding the performance of individual probes, it was evident that three of those that showed a limited sensitivity to detect minority genomes in binary mixtures of pure clones (S68, T69b and K70a) did produce hybridization signals with clinical samples including the complementary sequences at proportions of only 1.00–4.99% of the quasispecies (e.g., A20). Therefore, although the use of mixtures of clonal samples was useful to set a preliminary cut-off of 10%, it did not define the maximum sensitivity for the detection of minority subpopulations in HIV-1 quasispecies.

Discussion

HIV infection is usually diagnosed by testing serum for antibodies to HIV using a commercially available enzyme-linked immunosorbent assay (ELISA or EIA). Because the ELISA test is not entirely specific, positive results are confirmed with a Western blot assay, which identifies antibodies to specific components of HIV (Schwartz, Dans, and Kinosian, 1988). The 2-step process may mean that a patient must wait for a week or more to receive test results. ELISA is quite sensitive in chronic HIV infection (although decline in antibody responses have been reported in advanced AIDS), but because antibody production does not occur immediately upon infection, an infected individual may test ELISA negative during a "window period" that varies in length from a few weeks to a few months after infection, depending on the individual case and assay used. Despite negative antibody testing during this window period, an individual may have high viral load and be at high risk of transmitting infection.

Newer methodologies allow antibody testing on saliva (Emmons et al., 1995; Martínez et al., 1999) and urine (Martínez et al., 1999; Tiensiwakul, 1998) specimens, although positive results should be confirmed with serologic testing. Home testing methods are also available (Colfax et al., 2002). Rapid HIV serum testing, with results available in 3-30 minutes, has shown 99-100% sensitivity and specificity compared to ELISA when tested in clinical settings (Ketema et al., 2002), including in resource-poor settings (Phili and Vardas, 2002; Respess, Rayfield, and Dondero, 2001) and in pooled specimens (Soroka et al., 2003). In recent years, with the availability of rapid tests such as OraQuick (Abbott Laboratories, Abbott Park, IL) and Reveal (MedMira, Halifax, Nova Scotia, Canada), rapid testing protocols are being implemented in many countries, and will likely become commonplace.

A widely used variant of microarray hybridization involves the immobilization of DNA fragments representing the different genes (or their parts) of a target organism and the subsequent hybridization of the microarray with samples under study. While this approach is very efficient for gene identification and quantification of mRNA profiles in cells and tissues, it is not suitable for the detection of minor genetic differences (low genomic divergence or single point mutations) between closely related species.

New antiretroviral medications that are in development include improved formulations of currently approved drugs (to enhance bioavailability, increase half-life, or reduce adverse effects); new drugs in the same classes as currently approved drugs (such as PIs or NNRTIs with fewer adverse effects or unique resistance patterns); and drugs with novel mechanisms of action (eg, integrase inhibitors, entry inhibitors, and HIV co receptor blockers).

Several preliminary experiments showed a notorious decrease in the quality of results using aldehyde coated slides, streptavidine coated magnetic beads to obtain single-stranded DNA or a formamide hybridization solution. Additionally, other conditions involving nucleotide probes of different length, presence of absence of spacers between the array substrate and the probe, and different labelling and hybridization conditions were tested. The best signal to noise ratios and the most reproducible results were achieved using 15-mer with oligo $(dT)_{15}$ spacer or $(dCCT)_5$ spacer and 50µM concentrated oligonucleotide probes, with the queried position located towards the center of the probe, printed of super-epoxi-coated slides (experimental conditions detailed in Materials and Methods). Hybridization and washing temperatures were also selected after systematic preliminary experiments.

The stepwise filtering method developed has allowed discarding a number of individual spots and probes without compromising the number of complete arrays to be included in the analysis. Therefore, this procedure maximized the recovery of useful information from the experimental data. Indeed, before filtering, 9.7% of the microarrays hybridized with samples from the training set and 12.6% of those hybridized with clinical samples (8.4% of the PR region and , 4.2% of the RT) showed an excess of positive signals that could have recommend their elimination. Nevertheless, after filtering spots and probes, the amount of complete microarrays to be discarded decreased to 4.46% in the training set and 4.16% (PR arrays) or even 0% (RT arrays) in the test set (Table 3). In addition, it became evident that quality control filters must be applied sequentially instead of independently in order to get better classification accuracy. This is mainly due to the fact that each filter collaborates to the progressive decrease of the rate of FP signals. Thus, by the sequential use

of the four steps of the filtering protocol, the ratio of correctly classified data significantly increased in both the training and the test sets, concomitantly with a neat reduction of FP and FN signals. On the contrary, the negligible fraction of UD spots was not significantly affected by the stepwise filtering process.

In any case, after the stepwise filtering protocol FP signals remained the main source of errors in the training and the test sets. Most of the FPs is due to the fact that the theoretical hybridization tables have been defined in a very conservative way, assuming that a single mismatch between probe and target at positions different to those at the 5' or 3' end of the hybridizing sequence should always produce a negative signal. Nevertheless, this is not the case with most of the high throughput genotyping microarrays.

In turn, some of the FN signals are produced because most of the clinical samples of the test set have been obtained from multi-treated patients, and they show a number of mutant nucleotides at positions adjacent to the queried codon that prevent the correct hybridization to its corresponding probe. Additionally, it is important to take into account that the sequences covered by contiguous probes very often overlap, in such a way that a mutation in the sample can affect its hybridization to different probes. As an example, hybridization with plasmid pINS produced reduced signals in the probes D67, S68, T69 and K70.

The analysis performed is also useful for identifying individual probes that compromise the overall accuracy of the genotyping microarray, thus being clear candidates to be redesigned for further versions of it. As an example, probe Y181 (with an overlap between its positive and negative curves of 13.5%, see Supplementary Fig. S1) showed a high proportion of FN signals with samples from the training set. By discarding this probe in the final classification, we increase the percentage of correctly classified spots from 93.33% to 97.21% (remaining 1.77% of FP, 0.75% of FN and 0.26% of UD). Nevertheless, the inclusion of this probe in the test set did not change the overall performance of the genotyping microarray in the RT region (89.50% of correct, 6.56% of FP, 2.28% of FN and 1.65% of UD spots).

We have demonstrated that the genotype-specific oligonucleotides unambiguously identified different drug-resistance mutations in PR- and RT-HIV-1. The use of these oligonucleotides for solid-phase hybridization in a microarray format, with fluorescently labelled single strand DNA, proved to be an efficient tool for rapid genotyping. The simplicity of the proposed microarray protocol, together with its use of a large number of species-specific oligoprobes and its ability to analyse multiple samples in a short time, offers clear

advantages. This approach appears to be highly robust and informative and can become a versatile tool being adapted to analyse a broad variety of micro-organisms.

References

Allers, K., Knoepfel, S. A., Rauch, P., Walter, H., Opravil, M., Fischer, M., Gunthard, H. F., and Metzner, K. J. (2007). Persistence of lamivudine-sensitive HIV-1 quasispecies in the presence of lamivudine in vitro and in vivo. J Acquir Immune Defic Syndr **44**(4)**,** 377-85.

Amexis, G., Oeth, P., Abel, K., Ivshina, A., Pelloquin, F., Cantor, C. R., Brau, A., and Chumakov, K. (2001). Quantitative mutant analysis of viral quasispecies by chip-based matrix-assisted laser desorption/ ionization time-of-flight mass spectrometry. Proc. Natl. Acad. Sci. USA **98**(21)**,** 12097-102.

Balanant, J., Guillot, S., Candrea, A., Delpeyroux, F., and Crainic, R. (1991). The natural genomic variability of poliovirus analyzed by a restriction fragment length polymorphism assay. Virology **184**(2)**,** 645-54.

Barlow, K. L., Green, J., and Clewley, J. P. (2000). Viral genome characterisation by the heteroduplex mobility and heteroduplex tracking assays. Rev Med Virol **10**(5)**,** 321-35.

Bergroth, T., Sonnerborg, A., and Yun, Z. (2005). Discrimination of lamivudine resistant minor HIV-1 variants by selective real-time PCR. J Virol Methods **127**(1)**,** 100-7.

Bolstad, B. M., Irizarry R. A., Astrand, M, and Speed, T. P. (2003). A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics 19: 185-193.

Briones, C., Mas, A., Gomez-Mariano, G., Altisent, C., Menendez-Arias, L., Soriano, V., and Domingo, E. (2000). Dynamics of dominance of a dipeptide insertion in reverse transcriptase of HIV-1 from patients subjected to prolonged therapy. Virus Res **66**(1)**,** 13-26.

Burke, D. S. (1997). Recombination in HIV: an important viral evolutionary strategy. Emerg Infect Dis **3**(3)**,** 253-9.

Cai, F., Chen, H., Hicks, C. B., Bartlett, J. A., Zhu, J., and Gao, F. (2007). Detection of minor drug-resistant populations by parallel allele-specific sequencing. Nat Methods **4**(2)**,** 123-5.

Coffin, J. M. (1995). HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science **267**(5197)**,** 483-489.

Colfax, G. N., Lehman, J. S., Bindman, A. B., Vittinghoff, E., Vranizan, K., Fleming, P. L., Chesney, M., Osmond, D., and Hecht, F. M. (2002). What happened to home HIV test collection kits? Intent to use kits, actual use, and barriers to use among persons at risk for HIV infection. AIDS **14(5),** 675-682.

Charpentier, C., Dwyer, D. E., Mammano, F., Lecossier, D., Clavel, F., and Hance, A. J. (2004). Role of minority populations of human immunodeficiency virus type 1 in the evolution of viral resistance to protease inhibitors. J. Virol. **78**(8)**,** 4234-47.

Cherkasova, E., Laassri, M., Chizhikov, V., Korotkova, E., Dragunsky, E., Agol, V. I., and Chumakov, K. (2003). Microarray analysis of evolution of RNA viruses: evidence of circulation of virulent highly divergent vaccine-derived polioviruses. Proc. Natl. Acad. Sci. USA **100**(16)**,** 9398-403.

Chizhikov, V., Rasooly, A., Chumakov, K., and Levy, D. D. (2001). Microarray analysis of microbial virulence factors. Appl Environ Microbiol **67**(7)**,** 3258-63.

Chizhikov, V., Wagner, M., Ivshina, A., Hoshino, Y., Kapikian, A. Z., and Chumakov, K. (2002). Detection and genotyping of human group A rotaviruses by oligonucleotide microarray hybridization. J. Clin. Microbiol. **40**(7)**,** 2398-407.

Chumakov, K. M., Powers, L. B., Noonan, K. E., Roninson, I. B., and Levenbook, I. S. (1991). Correlation between amount of virus with altered nucleotide sequence and the monkey test for acceptability of oral poliovirus vaccine. Proc. Natl. Acad. Sci. USA **88**(1)**,** 199-203.

Domingo, E. (1992). Genetic variation and quasi-species. Curr. Opin. Genet. Dev **2**(1)**,** 61-63.

Domingo, E. (1999). RNA virus quasispecies as models of biological complexity. In "Proceedings 10th Anniversary Symposium. The Samuel Roberts Noble Foundation, Plant Biology Division" (R. A. Dixon, M. J. Harrison, and M. J. Roossinck, Eds.), pp. 79-90. The Samuel Roberts Foundation, Ardmore.

Domingo, E., and Holland, J. J. (1992). Complications of RNA heterogeneity for the engineering of virus vaccines and antiviral agents. Genet. Eng. (N Y) **14,** 13-31.

Doukhan, L., and Delwart, E. (2001). Population genetic analysis of the protease locus of human immunodeficiency virus type 1 quasispecies undergoing drug selection, using a denaturing gradient-heteroduplex tracking assay. J Virol **75**(14)**,** 6729-36.

Eigen, M., and Biebricher, C. K. (1988). Sequence space and quasispecies distribution. In "RNA Genetics" (E. Domingo, P. Ahlquist, and J. J. Holland, Eds.), Vol. 3, pp. 211-245. CRC Press, Boca Raton, FL.

Emmons, W. W., Paparello, S. F., Decker, C. F., Sheffield, J. M., and Lowe-Bey, F. H. (1995). A modified ELISA and western blot accurately determine anti-human immunodeficiency virus type 1 antibodies in oral fluids obtained with a special collecting device. J. Infect. Dis. **171(6),** 1406-1410.

Forgy, E.W. (1965). Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics 21: 768-769.

Gerry, N. P., Witowski, N. E., Day, J., Hammer, R. P., Barany, G., and Barany, F. (1999). Universal DNA microarray method for multiplex detection of low abundance point mutations. J. Mol. Biol. **292,** 251-262.

Grubbs, F.E. (1950). Sample Criteria for testing outlying observations. Ann. Math. Stat. 21: 27-58.

Gulick, R. M., Mellors, J. W., Havlir, D., Eron, J. J., Gonzalez, C., McMahon, D., Richman, D. D., Valentine, F. T., Jonas, L., Meibohm, A., Emini, E. A., and Chodakewitz, J. A. (1997). Treatment with indinavir, zidovudine, and lamivudine in adults with human immunodeficiency virus infection and prior antiretroviral therapy. N. Engl. J. Med. **337**(11)**,** 734-9.

Guo, Z., Guilfoyle, R. A., Thiel, A. J., Wang, R., and Smith, L. M. (1994). Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. Nucleic Acids Res. **22**(24)**,** 5456-65.

Hammer, S. M., Squires, K. E., Hughes, M. D., Grimes, J. M., Demeter, L. M., Currier, J. S., Eron, J. J., Jr., Feinberg, J. E., Balfour, H. H., Jr., Deyton, L. R., Chodakewitz, J. A., and Fischl, M. A. (1997). A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. AIDS Clinical Trials Group 320 Study Team. N. Engl. J. Med. **337**(11)**,** 725-33.

Hance, A. J., Lemiale, V., Izopet, J., Lecossier, D., Joly, V., Massip, P., Mammano, F., Descamps, D., Brun-Vezinet, F., and Clavel, F. (2001). Changes in human immunodeficiency virus type 1 populations after treatment interruption in patients failing antiretroviral therapy. J Virol **75**(14)**,** 6410-7.

Holland, J. J., de La Torre, J. C., and Steinhauer, D. A. (1992). RNA virus populations as quasispecies. Curr. Top. Microbiol. Immunol. **176,** 1-20.

Ivshina, A. V., Vodeiko, G. M., Kuznetsov, V. A., Volokhov, D., Taffs, R., Chizhikov, V. I., Levandowski, R. A., and Chumakov, K. M. (2004). Mapping of genomic segments of influenza B virus strains by an oligonucleotide microarray method. J Clin Microbiol **42**(12)**,** 5793-801.

Izopet, J., Souyris, C., Hance, A., Sandres-Saune, K., Alvarez, M., Pasquier, C., Clavel, F., Puel, J., and Massip, P. (2002). Evolution of human immunodeficiency virus type 1 populations after resumption of therapy following treatment interruption and shift in resistance genotype. J Infect Dis **185**(10)**,** 1506-10.

Kapoor, A., Jones, M., Shafer, R. W., Rhee, S. Y., Kazanjian, P., and Delwart, E. L. (2004). Sequencing-based detection of low-frequency human immunodeficiency virus type 1 drug-resistant mutants by an RNA/DNA heteroduplex generator-tracking assay. J Virol **78**(13)**,** 7112-23.

Ketema, F., Zeh, C., Edelman, D. C., Saville, R., and Constantine, N. T. (2002). Evaluation of a rapid human immunodeficiency virus test at two community clinics in Kwazulu-Natal. S Afr Med J **92(10),** 818-821.

Kijak, G. H., Simon, V., Balfe, P., Vanderhoeven, J., Pampuro, S. E., Zala, C., Ochoa, C., Cahn, P., Markowitz, M., and Salomon, H. (2002). Origin of human immunodeficiency virus type 1 quasispecies emerging after antiretroviral treatment interruption in patients with therapeutic failure. J. Virol. **76**(14)**,** 7000-9.

Lapa, S., Mikheev, M., Shchelkunov, S., Mikhailovich, V., Sobolev, A., Blinov, V., Babkin, I., Guskov, A., Sokunova, E., Zasedatelev, A., Sandakhchiev, L., and Mirzabekov, A. (2002). Species-level identification of orthopoxviruses with an oligonucleotide microchip. J Clin Microbiol **40**(3)**,** 753-7.

Lecossier, D., Shulman, N. S., Morand-Joubert, L., Shafer, R. W., Joly, V., Zolopa, A. R., Clavel, F., and Hance, A. J. (2005). Detection of minority populations of HIV-1 expressing the K103N resistance mutation in patients failing nevirapine. J Acquir Immune Defic Syndr **38**(1)**,** 37-42.

Mansky, L. M., and Temin, H. M. (1995). Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J. Virol. **69**(8)**,** 5087-94.

Martin, V., Perales, C., Abia, D., Ortiz, A. R., Domingo, E., and Briones, C. (2006). Microarray-based identification of antigenic variants of foot-and-mouth disease virus: a bioinformatics quality assessment. BMC Genomics **7,** 117.

Martínez, P. M., Torres, A. R., Ortiz de Lejarazu, R., Montoya, A., Martín, J. F., and Eiros, J. M. (1999). Human immunodeficiency virus antibody testing by enzyme-linked fluorescent and western blot assays using serum, gingival-crevicular transudate, and urine samples. J Clin Microbiol **37(4),** 1100-1106.

Menéndez-Arias, L. (2002). Targeting HIV: antiretroviral therapy and development of drug resistance. Trends in Pharmacological Science **23**(8)**,** 381-8.

Metzner, K. J., Bonhoeffer, S., Fischer, M., Karanicolas, R., Allers, K., Joos, B., Weber, R., Hirschel, B., Kostrikis, L. G., and Gunthard, H. F. (2003). Emergence of minor populations of human immunodeficiency virus type 1 carrying the M184V and L90M mutations in subjects undergoing structured treatment interruptions. J. Infect. Dis. **188**(10)**,** 1433-43.

Metzner, K. J., Rauch, P., Walter, H., Boesecke, C., Zollner, B., Jessen, H., Schewe, K., Fenske, S., Gellermann, H., and Stellbrink, H. J. (2005). Detection of minor populations of drug-resistant HIV-1 in acute seroconverters. Aids **19**(16)**,** 1819-25.

Palmer, S., Shafer, R. W., and Merigan, T. C. (1999). Highly drug-resistant HIV-1 clinical isolates are cross-resistant to many antiretroviral compounds in current clinical development. AIDS **13**(6)**,** 661-7.

Paolucci, S., Baldanti, F., Campanini, G., Zavattoni, M., Cattaneo, E., Dossena, L., and Gerna, G. (2001). Analysis of HIV drug-resistant quasispecies in plasma, peripheral blood mononuclear cells and viral isolates from treatment-naive and HAART patients. J Med Virol **65**(2)**,** 207-17.

Pawlotsky, J. M. (2006). Hepatitis C virus population dynamics during infection. Current Topics in Microbiol. and Immunol. **299,** 261-284.

Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., and Ho, D. D. (1996). HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science **271**(5255)**,** 1582-1586.

Phili, R., and Vardas, E. (2002). Evaluation of a rapid human immunodeficiency virus test at two community clinics in Kwazulu-Natal **92(10),** 818-821.

Resch, W., Parkin, N., Stuelke, E. L., Watkins, T., and Swanstrom, R. (2001). A multiple-site-specific heteroduplex tracking assay as a tool for the study of viral population dynamics. Proc Natl Acad Sci U S A **98**(1)**,** 176-81.

Respess, R. A., Rayfield, M. A., and Dondero, T. J. (2001). Laboratory testing and rapid HIV assays: applications for HIV surveillance in hard-to-reach populations. AIDS **15(3),** S49-59.

Schmit, J. C., Cogniaux, J., Hermans, P., Van Vaeck, C., Sprecher, S., Van Remoortel, B., Witvrouw, M., Balzarini, J., Desmyter, J., De Clercq, E., and Vandamme, A. M. (1996). Multiple drug resistance to nucleoside analogues and nonnucleoside reverse transcriptase inhibitors in an efficiently replicating human immunodeficiency virus type 1 patient strain. J. Infect. Dis. **174**(5)**,** 962-8.

Schmit, J. C., Van Laethem, K., Ruiz, L., Hermans, P., Sprecher, S., Sonnerborg, A., Leal, M., Harrer, T., Clotet, B., Arendt, V., Lissen, E., Witvrouw, M., Desmyter, J., De Clercq, E., and Vandamme, A. M. (1998). Multiple dideoxynucleoside analogue-resistant (MddNR) HIV-1 strains isolated from patients from different European countries. AIDS **12**(15)**,** 2007-15.

Schwartz, J. S., Dans, P. E., and Kinosian, B. P. (1988). Human immunodeficiency virus test evaluation, performance, and use. Proposals to make good tests better. JAMA **259(17),** 2574-2579.

Soroka, S. D., Granade, T. C., Phillips, S., and Parekh, B. (2003). The use of simple, rapid tests to detect antibodies to human immunodeficiency virus types 1 and 2 in pooled serum specimens. J Clin Virol **27(1),** 90-96.

Tiensiwakul, P. (1998). Urinary HIV-1 antibody patterns by western blot assay. Clin Lab Sci.

Volokhov, D., Rasooly, A., Chumakov, K., and Chizhikov, V. (2002). Identification of Listeria species by microarray-based assay. J Clin Microbiol **40**(12)**,** 4720-8.

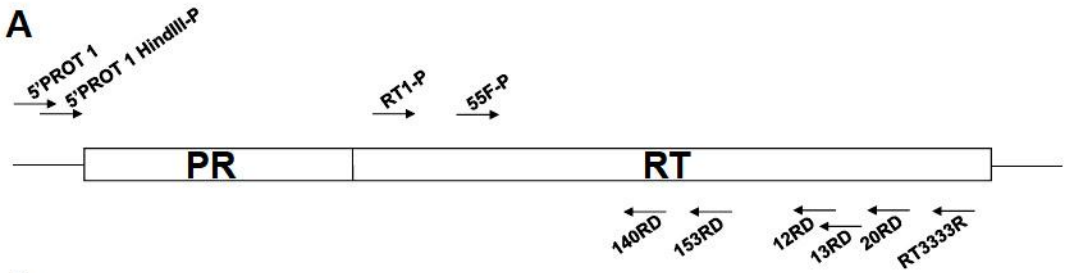
Wang, D., Coscoy, L., Zylberberg, M., Avila, P. C., Boushey, H. A., Ganem, D., and DeRisi, J. L. (2002). Microarray-based detection and genotyping of viral pathogens. Proc. Natl. Acad. Sci. USA **99**(24)**,** 15687-92.

Westby, M., Lewis, M., Whitcomb, J., Youle, M., Pozniak, A. L., James, I. T., Jenkins, T. M., Perros, M., and van der Ryst, E. (2006). Emergence of CXCR4-using human immunodeficiency virus type 1 (HIV-1) variants in a minority of HIV-1-infected patients following treatment with the CCR5 antagonist maraviroc is from a pretreatment CXCR4-using virus reservoir. J Virol **80**(10)**,** 4909-20.

Wodarz, D., and Nowak, M. A. (1999). Dynamics of HIV pathogenesis and treatment. In "Origin and evolution of viruses" (E. Domingo, R. G. Webster, and J. J. Holland, Eds.), pp. 197-223. Academic Press, San Diego.

Wong, J. K., Gunthard, H. F., Havlir, D. V., Zhang, Z. Q., Haase, A. T., Ignacio, C. C., Kwok, S., Emini, E., and Richman, D. D. (1997). Reduction of HIV-1 in blood and lymph nodes following potent antiretroviral therapy and the virologic correlates of treatment failure. Proc. Natl. Acad. Sci. USA **94**(23)**,** 12574-9.

FIGURES



B

Primer Names	SEQUENCE (5'-3')
5' PROT1	AGGCTAATTTTTTAGGGAAAATCTGGCCTTCC
5' PROT HindIII-P	TCAGAGCAGACCAGAGCCAACAGCCCCACC
RT1-P	CCAAAAGTTAAACAATGGCCATTG
55F-P	CAAAAATTGGGCCTGAAAATCC
140RD	CCAGGGATTAGATATCAGTACAATG
153RD	GGAAAGGATCACCAGCAATA
12RD	AACATCAGAAAGAACCTCCA
13RD	CCTTTGGATGGGTTATGAAC
20RD	GCCAGAAAAGGACAGCTGGACTGT
RT3333R	CAATGACATACAGAAGTTAGTGGG

Fig.1

Figure 1.- Scheme of oligonucleotides used for the HIV-PR and HIV-RT amplification. A). The genomic fragment representing genes, protease (PR) and retrotranscriptase (RT) of human immunodeficiency virus was amplified by different pairs of oligonucleotides showed in the scheme as rows depicted approximately in their corresponding genomic sequence position. B) Table of primer sequences used in amplification.

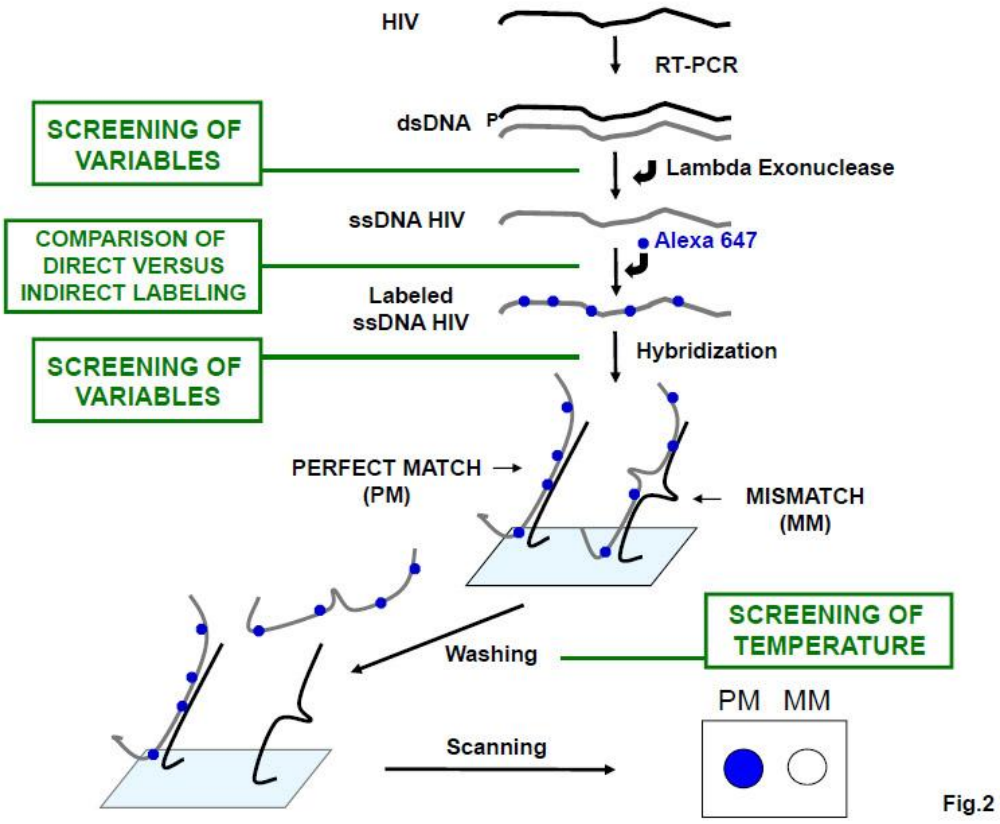


Figure 2.- Scheme of the successive steps from the copying of HIV genomic RNA to scanning of the microarray. PCR was performed using different primers (Figure 1). Green boxes indicate those steps for which a number of variables were tested. The final protocol used for the different steps is detailed in Materials and Methods.

VIROCHIP 4.0

Oligo Names	**SEQUENCE (5'-3')**
ICH-PR3	TTTTTTTTTTTTTTTGATACAGGAGCAGAT
ICH-PR4	TTTTTTTTTTTTTTTGTTGCACTTTAAATTTT
R8-PR-2	TTTTTTTTTTTTTTTGCAACGACCCCTC
Q8-PR-2	TTTTTTTTTTTTTTTGCAAC**A**ACCCCTC
D30-PR	TTTTTTTTTTTTTTTAGCAGATGATACAGT
N30-PR	TTTTTTTTTTTTTTTAGCAGAT**A**ATACAGT

M36-PR	TTTTTTTTTTTTTTTAAGAAATGAGTTTGC
I36-PR	TTTTTTTTTTTTTTTAAGAAAT**A**AGTTTGC
M46b-PR	TTTTTTTTTTTTTTTCAAAAATGATAGGGG
M46c-PR	TTTTTTTTTTTTTTTACCAAAAATGATAGG
F46-PR	TTTTTTTTTTTTTTTCCAAAA**T**TCATAGGG
I46-PR	TTTTTTTTTTTTTTTCAAAAAT**A**ATAGGGG
L46-PR	TTTTTTTTTTTTTTTACCAAAA**T**TGATAGG
V46-PR	TTTTTTTTTTTTTTTACCAAAA**G**TGATAGG
G48-PR	TTTTTTTTTTTTTTTATGATAGGGGGAATT
V48-PR	TTTTTTTTTTTTTTTATGATAG**T**GGGAATT
I50-PR	TTTTTTTTTTTTTTTGGGGAATTGGAG
L50a-PR	TTTTTTTTTTTTTTTGGGGA**T**TAGGAG
L50b-PR	TTTTTTTTTTTTTTTGGGA**C**TCGGAG
V50-PR-2	TTTTTTTTTTTTTTTGGTGA**G**TTGGAG
I54-PR	TTTTTTTTTTTTTTTAGGTTTTATCAAAGTA
V54-PR-2	TTTTTTTTTTTTTTTAGGTATT**G**TCAAAGTA
A71-PR-2	TCCTCCTCCTCCTCCCATAAAGCTATAGGTA
V71-PR-2	TCCTCCTCCTCCTCCCATAAAG**T**TATAGGTA
V82b-PR-2	TCCTCCTCCTCCTCCTACACCTGTCAACAT
A82-PR-2	TCCTCCTCCTCCTCCACACCTG**C**CAACAT
F82-PR-2	TCCTCCTCCTCCTCCTACACC**T**TTCAACAT
I82-PR-2	TCCTCCTCCTCCTCCTACACCT**A**TCAACAT
S82-PR-2	TCCTCCTCCTCCTCCTACACCT**TC**CAACAT
T82-PR-2	TCCTCCTCCTCCTCCTACACCT**AC**CAACAT
I84-PR-2	TCCTCCTCCTCCTCCTGTCAACATAATTGG
A84-PR-2	TCCTCCTCCTCCTCCGTCAAC**GC**AATTGG
V84-PR-2	TCCTCCTCCTCCTCCGTCAAC**G**TAATTGG

L90a-PR	TTTTTTTTTTTTTTTAATCTGTTGACTCAG
L90a-PR-2	TTTTTTTTTTTTTTTAATCTGTTGACTCAGA
I90-PR	TTTTTTTTTTTTTTTAATCTG**A**T**A**ACTCAG
I90-PR-2	TTTTTTTTTTTTTTTAATCTG**A**T**A**ACTCAGA
M90-PR	TTTTTTTTTTTTTTTAAATCTG**A**TGACTCA
M90-PR-2	TTTTTTTTTTTTTTTAATCTG**A**TGACTCAGA
FMDV-G142-15r	TTTTTTTTTTTTTTTCAAATCCCCGCGTGC
FMDV-E142-15r	TTTTTTTTTTTTTTTCAAATCCTCGCGTGC
H2O+ss	
ICH-RT3	TTTTTTTTTTTTTTTATGGCCATTGACAGA
ICH-RT4	TTTTTTTTTTTTTTTATCTTAGAGCCTTTTA
M41	TTTTTTTTTTTTTTTTACAGAGATGGAAAA
L41a	TTTTTTTTTTTTTTTTACAGAG**T**TGGAAAA
L41b	TTTTTTTTTTTTTTTTACAGAG**C**TGGAAAA
A62	TTTTTTTTTTTTTTTAGTATTTGCCATAAAG
V62	TTTTTTTTTTTTTTTAGTATTTG**T**CATAAAG
K65-2	TCCTCCTCCTCCTCCCATAAAGAAAAAAGAC
R65-3	TCCTCCTCCTCCTCCCATAAAGA**G**AATAGAC
D67a-2	TCCTCCTCCTCCTCCAAAAAAGACAGTACTA
D67b-2	TCCTCCTCCTCCTCCGAAAAAAGACAGTACT
E67-2	TCCTCCTCCTCCTCCAAAAAAGA**A**AGTACTAA
G67	TTTTTTTTTTTTTTTAAAAAAG**G**CAGTACTA
G67-2	TCCTCCTCCTCCTCCAAAAAAG**G**CAGTACTA
N67-2	TCCTCCTCCTCCTCCGAAAAAAA**A**ACAGTACT
Del67	TTTTTTTTTTTTTTTAAGAAAA**A**A**A**GTACTAA
S68	TTTTTTTTTTTTTTTAAAAGACAGTACTAAAT

N68	TTTTTTTTTTTTTTTAAAAGACA**A**TACTAAAT
T69b	TTTTTTTTTTTTTTTGACAGTACTAAATGG
T69c	TTTTTTTTTTTTTTTGACAGTAC**A**AAATGG
A69	TTTTTTTTTTTTTTTAGACAGT**G**CTAAATG
D69	TTTTTTTTTTTTTTTGACAGT**GA**TAAATGG
G69	TTTTTTTTTTTTTTTGACAGT**GG**TAAATG
S69a	TTTTTTTTTTTTTTTGACAGTA**G**TAAATGG
S69b	TTTTTTTTTTTTTTTGACAGTA**GA**AAATGG
S69R70	TTTTTTTTTTTTTTTACAGT**T**CTAGATGGA
Ins69a	TTTTTTTTTTTTTTTA**G**T**AGTAGT**AGTAAAT
Ins69b	TTTTTTTTTTTTTTTGT**T**CT**AGTTCT**AGAT
Ins69c	TTTTTTTTTTTTTTTT**CG**AG**T**AGTTCGAAA
Ins69d	TTTTTTTTTTTTTTTAGTA**G**T**AGTGGT**AAA
Ins69e	TTTTTTTTTTTTTTTTAGT**GG**T**AGATGG**
Ins69f	TTTTTTTTTTTTTTTAGTA**G**T**AGCGCT**AAA
Ins69g	TTTTTTTTTTTTTTTACT**AGCAGCGCTA**
Ins69h	TTTTTTTTTTTTTTTAGTA**G**T**GAAGCA**AAA
Ins69i	TTTTTTTTTTTTTTTAGT**T**CT**ACCTCT**A**G**A
Ins69j	TTTTTTTTTTTTTTTAGTA**GCGTGACT**AAA
Ins69k	TTTTTTTTTTTTTTTAGTACT**AGTDST**AAAT
Ins69l	TTTTTTTTTTTTTTTGT**WS**T**AGTDST**A**R**AT
K70a	TTTTTTTTTTTTTTTCAGTACTAAATGGAG
K70b	TTTTTTTTTTTTTTTGTACTAAATGGAGAA
E70	TTTTTTTTTTTTTTTCAGTACT**G**AATGGA
N70a	TTTTTTTTTTTTTTTGTACTAA**T**TGGAGAA
N70b	TTTTTTTTTTTTTTTGTACTAA**C**TGGAGAA
R70a	TTTTTTTTTTTTTTTAGTACTA**G**ATGGAGA

R70b	TTTTTTTTTTTTTTTGTACTA**GG**TGGAG
L74	TTTTTTTTTTTTTTTGAGAAAATTAGTAGAT
V74-2	TTTTTTTTTTTTTTTGAGAATA**G**TAGTAGAT
V75	TTTTTTTTTTTTTTTGAAAATTAGTAGATTTC
I75	TTTTTTTTTTTTTTTGAAAATTA**A**TAGATTTC
T75	TTTTTTTTTTTTTTTGAAAATTA**AC**AGATTTC
F77	TTTTTTTTTTTTTTTAGTAGATTTCAGAGA
L77	TTTTTTTTTTTTTTTAGTAGAT**C**TCAGAGA
L100-2	TCCTCCTCCTCCTCCCGCAGGGTTAAAAAA
I100-2	TCCTCCTCCTCCTCCCGCAGGG**A**TAAAAAA
K101-3	TCCTCCTCCTCCTCCGCAGGGTTAAAAAAGA
E101-3	TCCTCCTCCTCCTCCGCAGGTTTA**G**AAAAGA
K103a-2	TCCTCCTCCTCCTCCAAAAAGAAAAAATCAGT
K103c-2	TCCTCCTCCTCCTCCAAAAAAGAAAAAATCAG
N103-3	TCCTCCTCCTCCTCCAAATAGAA**C**AAATCAGT
R103-3	TCCTCCTCCTCCTCCAAAATAGA**G**AAAATCAG
V106-2	TCCTCCTCCTCCTCCAAAATCAGTAACAGTA
A106-2	TCCTCCTCCTCCTCCAAAATCAG**C**AACAGTA
I106-2	TCCTCCTCCTCCTCCAAAAATCA**A**TAACAGTA
L106-2	TCCTCCTCCTCCTCCAAAAATCA**T**TAACAGTA
V108-2	TTTTTTTTTTTTTTTCAGTAACAGTACTGG
I108-2	TTTTTTTTTTTTTTTCAGTAACA**A**TACTGG
F116	TTTTTTTTTTTTTTTGCATATTTTTCAGTTC
Y116-3	TTTTTTTTTTTTTTTGCATATT**A**TTCACTTC
Q151-2	TTTTTTTTTTTTTTTGCTTCCACAGGGAT
M151-2	TTTTTTTTTTTTTTTGCTTCCA**AT**GGGAT
I178	TTTTTTTTTTTTTTTCAGACATAGTTATCT

M178	TTTTTTTTTTTTTTTCAGACAT**G**GTTATCT
V179-2	TTTTTTTTTTTTTTTAGACATAGTTATCTATC
D179-2	TTTTTTTTTTTTTTTAGACATAG**A**TATCTATC
E179-2	TTTTTTTTTTTTTTTAGACATAG**AG**ATCTAT
Y181	TTTTTTTTTTTTTTTAGTTATCTATCAATAC
C181	TTTTTTTTTTTTTTTAGTTATCT**G**TCAATAC
H181	TTTTTTTTTTTTTTTAGTTATC**C**ATCAATAC
I181	TTTTTTTTTTTTTTTAGTTATC**AT**TCAATAC
L181	TTTTTTTTTTTTTTTAGTTATC**CT**TCAATAC
M184a	TTTTTTTTTTTTTTTCAATACATGGATGATT
M184b	TTTTTTTTTTTTTTTCAATACATGGATGAT
I184	TTTTTTTTTTTTTTTCAATACAT**A**GATGATT
T184	TTTTTTTTTTTTTTTCAATACA**C**GGATGAT
V184b	TTTTTTTTTTTTTTTCAATAC**G**T**A**GATGAT
Y188a	TTTTTTTTTTTTTTTGATTTGTATGTAGGA
Y188c	TTTTTTTTTTTTTTTGATTTGTATGTAGGAT
C188-2	TTTTTTTTTTTTTTTGATTAGT**G**TGTAGGA
H188	TTTTTTTTTTTTTTTGATTTG**C**ATGTAGG
L188a	TTTTTTTTTTTTTTTGATTTGT**TA**GTAGGAT
L188b	TTTTTTTTTTTTTTTGATTTG**CT**TGTAGG
G190	TTTTTTTTTTTTTTTTATGTAGGATCTGAC
A190	TTTTTTTTTTTTTTTTATGTAG**C**ATCTGAC
E190	TTTTTTTTTTTTTTTTATGTAG**A**ATCTGAC
Q190	TTTTTTTTTTTTTTTTATGTA**CA**ATCTGAC
S190	TTTTTTTTTTTTTTTTATGTA**A**G**C**TCTGAC
T190	TTTTTTTTTTTTTTTTATGTAACATCTGAC
L210-3	TTTTTTTTTTTTTTTTATCTGTTGAGTTGGG

W210-2	TTTTTTTTTTTTTTTATCTGT**G**GAGTTGG
R211-2	TTTTTTTTTTTTTTTGTTGAGGTGGGGA
K211-2	TTTTTTTTTTTTTTTGTTGA**A**GTGGGGA
T215a-2	TTTTTTTTTTTTTTTGGACTTACCACACC
T215b-2	TTTTTTTTTTTTTTTGGGACTTACCACAC
C215-2	TTTTTTTTTTTTTTTGGACTT**TG**CACACC
F215-2	TTTTTTTTTTTTTTTGGACTT**TT**CACACC
S215-2	TTTTTTTTTTTTTTTGGGACTT**T**CCACAC
Y215-2	TTTTTTTTTTTTTTTGGACTT**TA**CACACC
K219	TTTTTTTTTTTTTTTACCAGACAAAAAACA
K219-2	TCCTCCTCCTCCTCCACCAGACAAAAAACA
E219	TTTTTTTTTTTTTTTACCAGAC**G**AAAAACA
Q219	TTTTTTTTTTTTTTTACCAGAC**C**AAAAACA
P225-2	TCCTCCTCCTCCTCCAAAGAACCTCCATTC
H225-2	TCCTCCTCCTCCTCCAAAGAAC**A**TCCATTC
M230-3	TTTTTTTTTTTTTTTCTTTGGATGGGTTAT
L230-3	TTTTTTTTTTTTTTTCTTTGG**C**TGGGTTAT
P236-2	TCCTCCTCCTCCTCCCTCCATCCTGATAAAT
L236-2	TCCTCCTCCTCCTCCCTCCATC**T**TGATAAAT
K238	TTTTTTTTTTTTTTTCCTGATAAATGGAC
T238	TTTTTTTTTTTTTTTCCTGATA**C**ATGGAC

Figure 3.- Oligonucleotide sequences printed on the microarray for the screening of HIV-PR and HIV-RT escape mutations after HIV treatments in infected patients. The column on the left shows the names of the oligonucleotides used in this work. The yellow backgrounds represent oligonucleotide sequences identical to the wild type nucleotide sequence. The white backgrounds represent oligonucleotides with a sequence corresponding to the different mutations tested. Green and grey names represent sequences used as negative hybridization controls and red names represent oligonucleotides used as positive controls. The nucleotide changes versus wild type sequence are specified in bold letter. The enquired position is located

at the centre of the oligonucleotide. The column on the right gives the predicted Tm value for each oligonucleotide, calculated according to Tm= 69.5+0.41 x (X% G+C)-650/total nucleotide number.

A

	A	B	C	D	E	F	G	H	I	J	K	L
1	CIH-PR3	CIH-PR4	R8-PR-2	Q8-PR-2	D30-PR	N30-PR	M36-PR	I36-PR	M46b-PR	M46c-PR	F46-PR	I46-PR
2	L46-PR	V46-PR	G48-PR	V48-PR	I50-PR	L50a-PR	L50b-PR	V50-PR-2	I54-PR	V54PR-2	A71-PR-2	V71-PR-2
3	V82b-PR-2	A82-PR-2	F82-PR-2	I82-PR-2	S82-PR-2	T82-PR-2	I84-PR-2	A84-PR-2	V84-PR-2	L90a-PR	L90a-PR-2	I50-PR
4	I50-PR-2	M90-PR	M90-PR-2	FMDV-G142-15r	FMDV-E142-15r	H_2O+SS	H_2O+SS	H_2O+SS	H_2O+SS	H_2O+SS	CIH-PR3	CIH-PR4
5	CIH-RT3	CIH-RT4	M41	L41a	L41b	A62	V62	K65-2	R65-3	D67a-2	D67b-2	E67-2
6	G67	G67-2	N67-2	Del67	S68	N68	T69b	T69c	A69	D69	G69	S69a
7	S69b	S69R70	Ins69a	Ins69b	Ins69c	Ins69d	Ins69e	Ins69f	Ins69g	Ins69h	Ins69i	Ins69j
8	Ins69k	Ins69l	K70a	K70b	E70	N70a	N70b	R70a	R70b	L74	V74-2	V75
9	I75	T75	F77	L77	L100-2	I100-2	K101-3	E101-3	K103a-2	K103c-2	N103-3	R103-3
10	V106-2	A106-2	I106-2	L106-2	V108-2	I108-2	F-116	Y116-3	Q151-2	M151-2	I178	M178
11	V179-2	D179-2	E179-2	Y181	C181	H181	I181	L181	M184a	M184b	I184	T184
12	V184b	Y188a	Y188c	C188-2	H188	L188a	L188b	G190	A190	E190	Q190	S190
13	T190	L210-3	W210-2	R211-2	K211-2	T215a-2	T215b-2	C215-2	F215-2	S215-2	Y215-2	K219
14	K219-2	E219	Q219	P225-2	H225-2	M230-3	L230-3	P236-2	L236-2	K238	T238	FMDV-G142-15r
15	FMDV-E142-15r	H_2O+SS	H_2O+SS	H_2O+SS	H_2O+SS	H_2O+SS	H_2O+SS	H_2O+SS	H_2O+SS	H_2O+SS	CIH-RT3	CIH-RT4

Figure 4.- Display of the oligonucleotides printed on the microarray for the screening of HIV-PR and HIV-RT escape mutations. Microarray hybridization patterns of wild type HIV-PR and HIV-RT. A). One hundred and fifty one oligonucleotides (50µM) were spotted in duplicate in each box in the scheme of the microarray depicted. Each box in the grid includes the name of each oligonucleotide in blue (sequence given in Figure 3) or the name of negative controls

written in green or gray (FMDV-G_{142}-15r, FMDV-E_{142}-15r, H_2O-ss). The names written in red in soft yellow background belong to the four oligonucleotides used as positive controls with highly conserved sequences in both different regions of HIV-PR and –RT (ICH-PR and ICH-RT). Oligonucleotides were distributed in 15 rows (4 for PR and 11 for RT) and 12 columns. The strong yellow boxes were oligonucleotides with wild type positions in the genomic fragment of the HIV-PR and HIV-RT. The oligonucleotides containing mutated positions were written in blue with white background. The pattern was printed four times in each microarray.

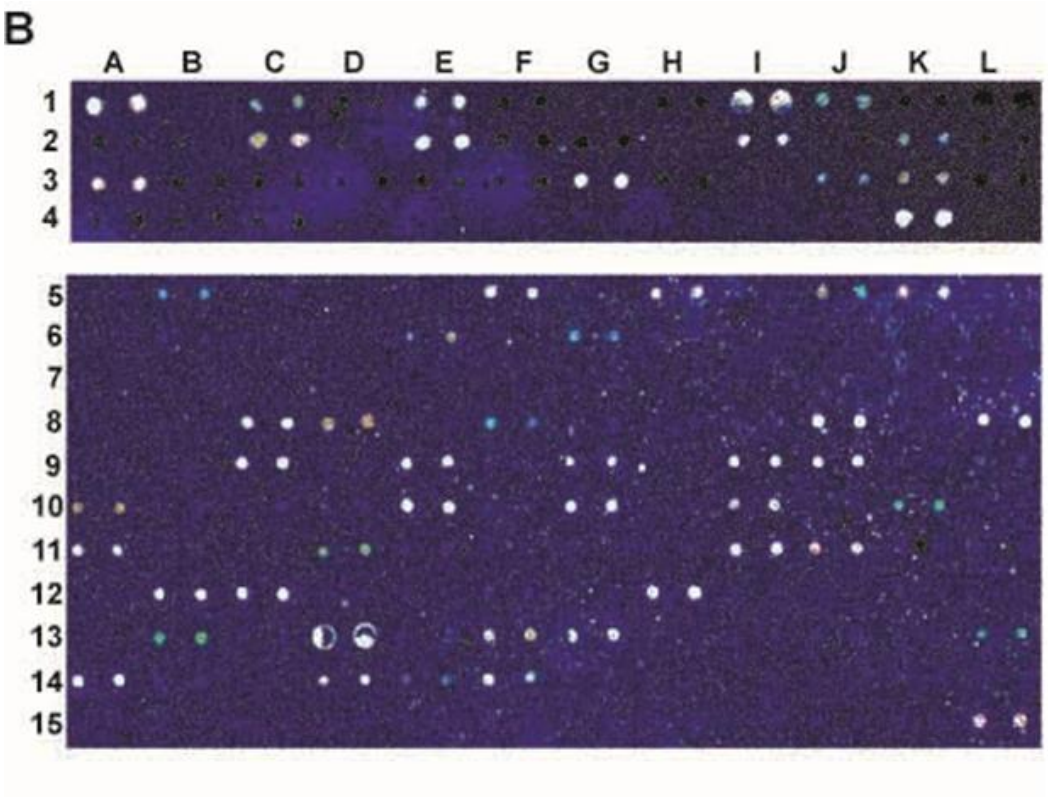


Figure 4B) The panels represent microarray images, given by the Alexa Fluor 647 fluorescence signal, after target hybridization, washing and scanning, as detailed in Materials and Methods. The distribution of oligonucleotide probes in each array is identical to that given in A. The target used in the top panel (four first rows) was a wild type HIV-PR and in the bottom panel a wild type HIV-RT, amplified from a plasmid where they were previously cloned. Positions expected to give a positive signal (perfect match) are those corresponding to strong yellow boxes in A. Due to the primers used to amplification the amino acid positions 236 and 238 are not detected.

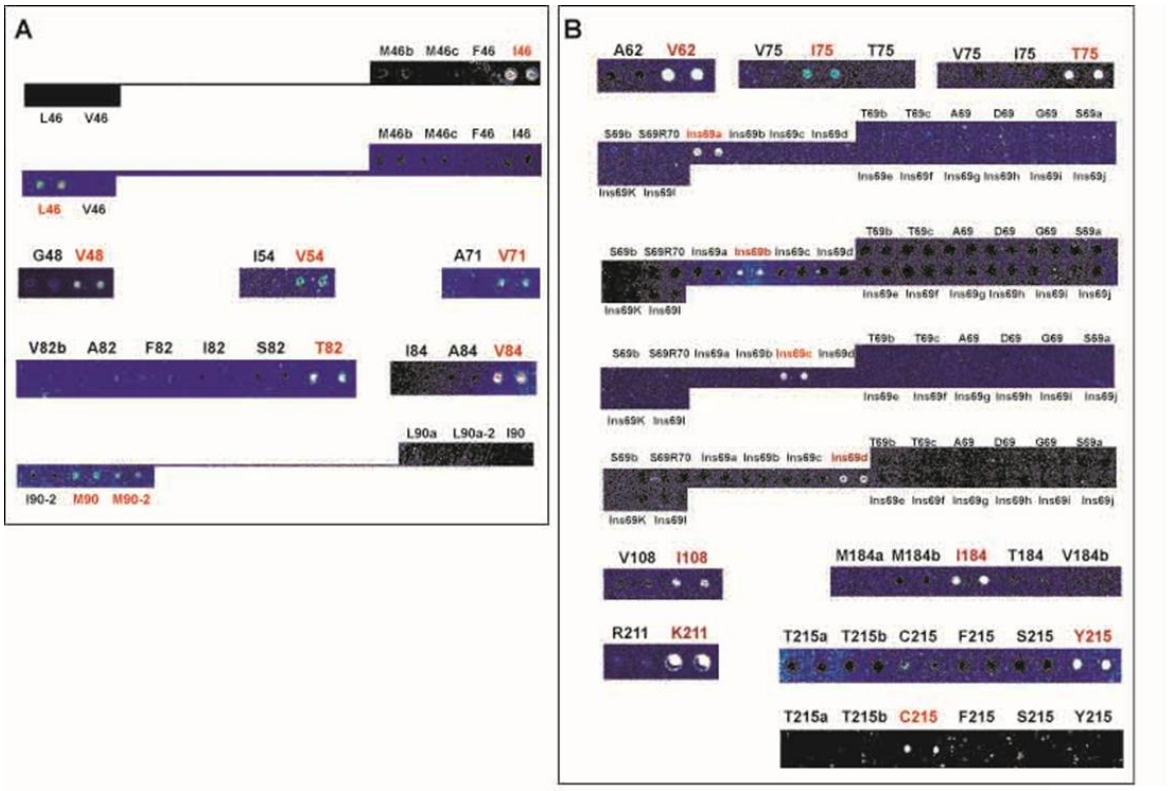


Figure 5.- Microarray hybridization patterns of HIV-PR and HIV-RT escape mutants cloned in a plasmid. Each panel represents a different position of **(A)** PRs or **(B)** RTs targets obtained from diverse microarray images, given by the Alexa 647 fluorescence signal, after hybridization, washing and scanning, as detailed in Materials and Methods. In red colour was written the mutant target tested, position expected to give a positive signal (perfect match), and in black the rest of the oligonucleotides, wild type or no, printed in the array corresponding to this position, expected to give a negative signal.

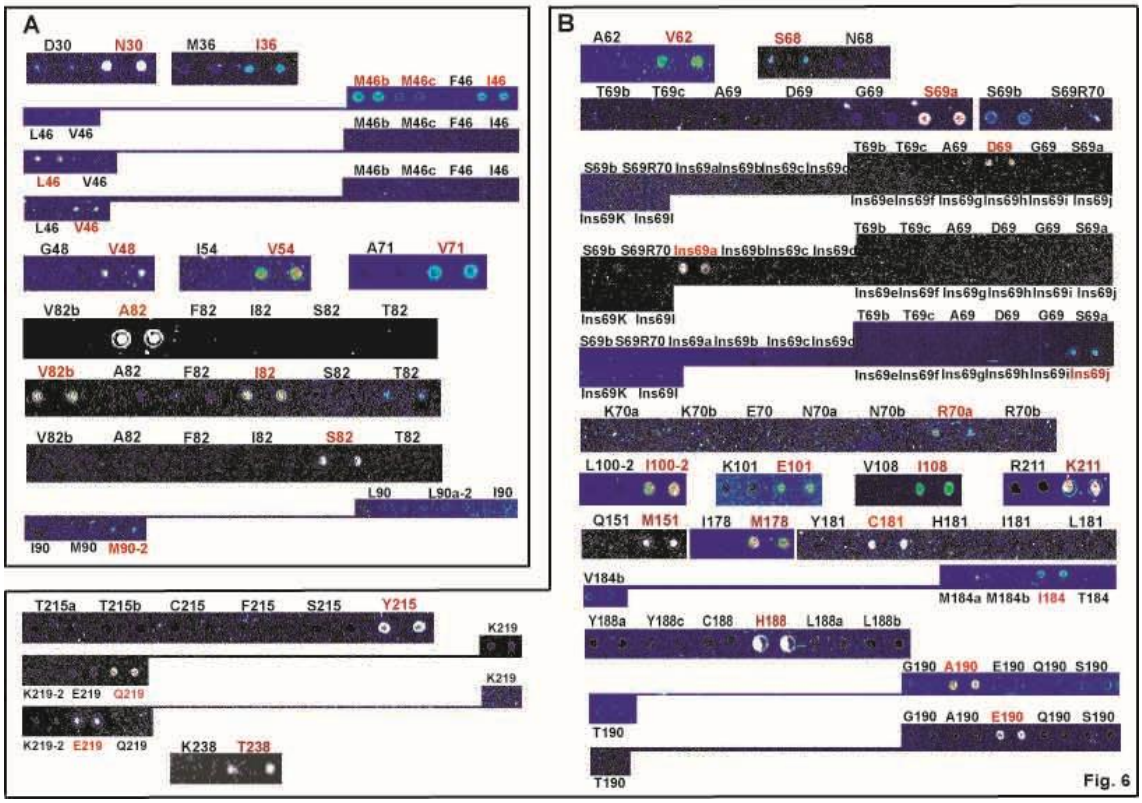


Figure 6.- Microarray hybridization patterns of HIV-PR and HIV-RT obtained from infected patient samples. Each panel represents different positions found mutated in diverse microarray images, given by the Alexa 647 fluorescence signal, after hybridization of different PRs (A) and RTs (B) targets amplified from HIV infected patients, washing and scanning, as detailed in Materials and Methods. In red colour was written the genotype target detected, position showing positive signal (perfect match), and in black the rest of oligonucleotides, wild type or no, printed in the array corresponding to this position, negative signal.

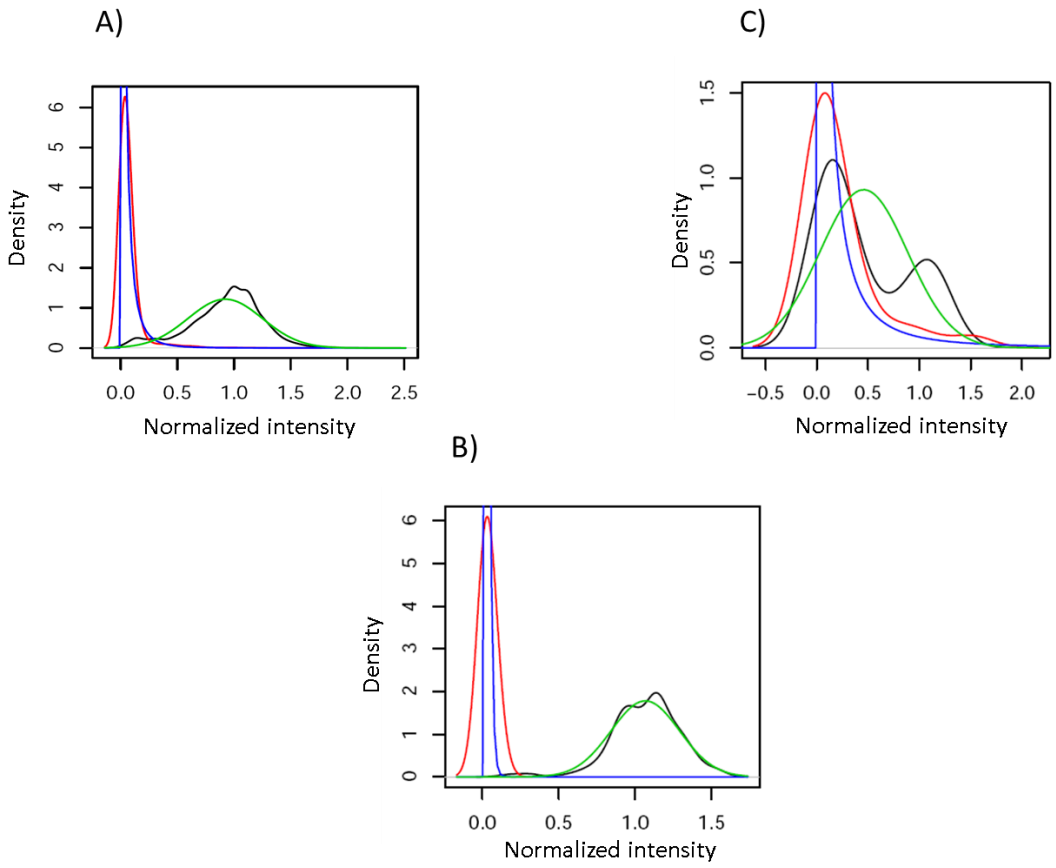


Figure 7.- Examples of calibrated curves for positive and negative signals from the training set. A) Global reference curves; **B)** Individual curves for probe Y188a, showing no overlapping between positive and negative curves. **C)** Individual curves for probe M230-3, with high overlapping (this probe was discarded during quality control). Legend: Red, density of normalized negative data; Blue, fit to a Log-normal distribution (negative signal); Black, density of normalized positive data; Green, fit to a Normal distribution (positive signal).

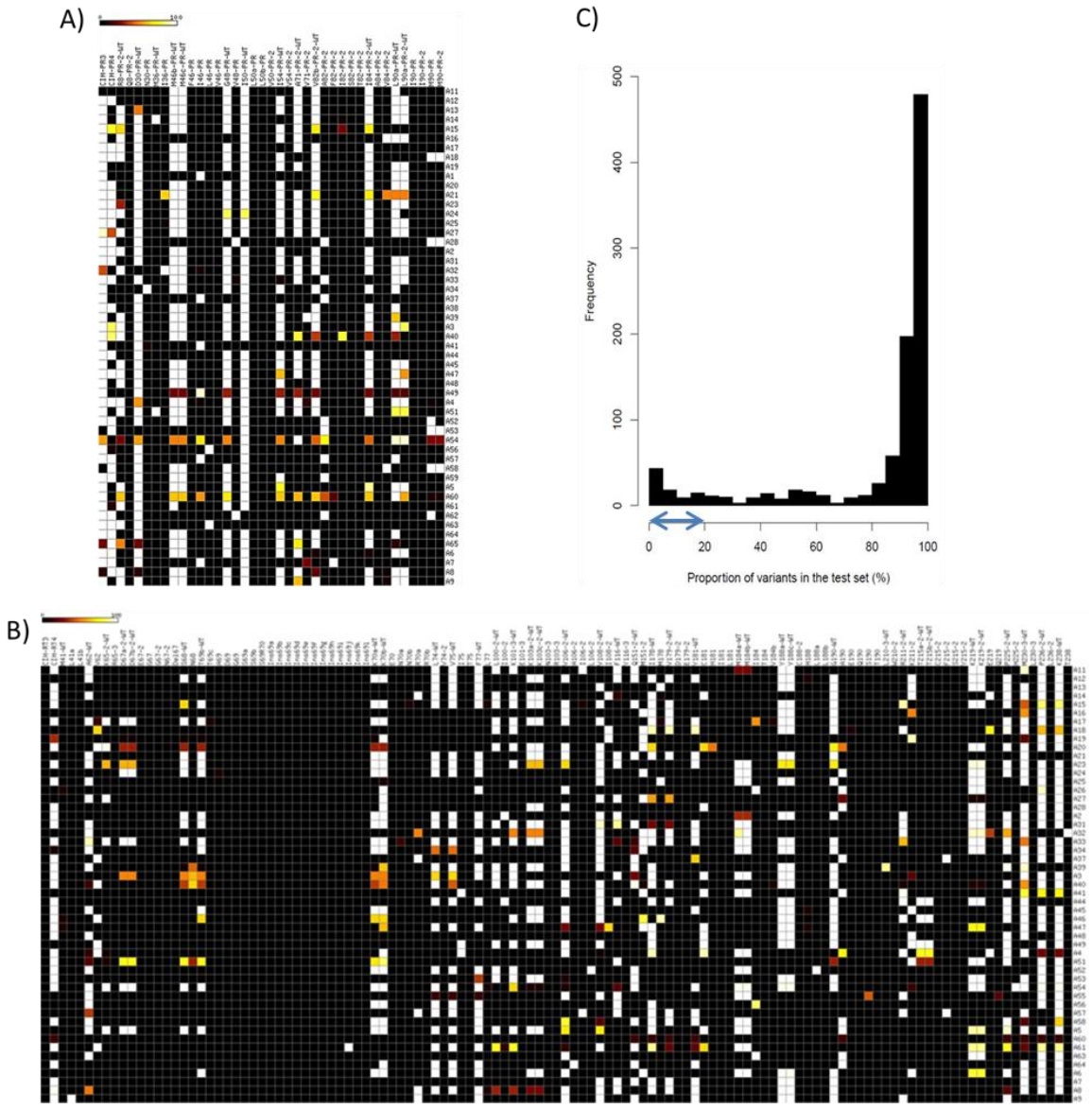


Figure 8.- Proportion of variants present in the clinical samples, determined by clonal sequencing. A) Codons queried by probes of the PR region (columns) vs. sequences of the target samples (rows); **B)** Codons queried by probes of the RT region; **C)** Distribution of variant proportions. The blue arrow indicates the region of minority variants.

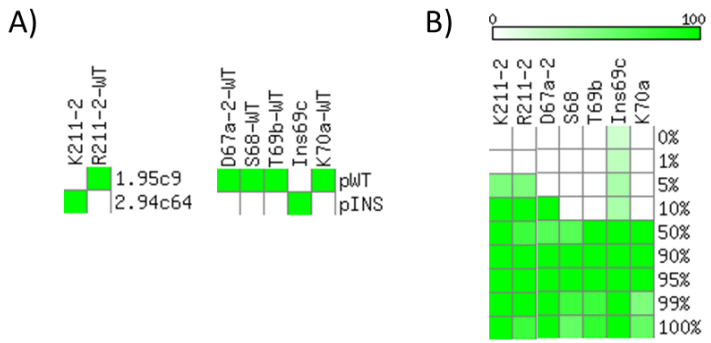


Figure 9.- Sensitivity of detection, estimated with binary mixtures of clonal samples. A) Theoretical hybridization tables of the pure samples used in the mixtures: 1.95c9/2.94c64 and pWT/pINS. Legend: green, expected positive hybridization; white, expected negative hybridization. **B)** Rate at which each sample in the mixture produces a positive hybridization at each probe identified in panel A. Bar shows the transition from white (positive signal not detected) to green (positive signal detected in every hybridization experiment).

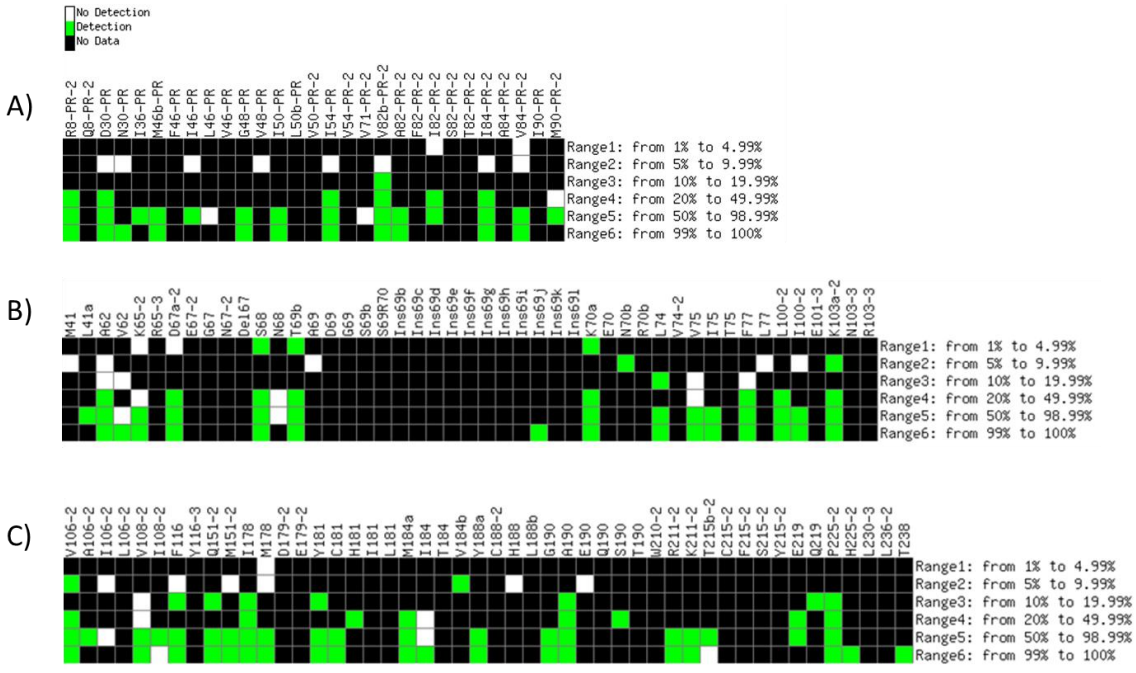


Figure 10.- Detection of minority variants in clinical samples by microarray analysis. Legend: green, detection at the corresponding probe when the complementary target is present within a given proportion range in the quasispecies; white, no detection; black, no data (the sequence complementary to the queried codon is not present in any of the clinical samples).

TABLES

Sample	Clinical genotype (consensus sequence)		Microarray genotype		% Molecular cloning	
	PR	**RT**	**PR**	**RT**	**PR**	**RT**
A1	-	-	-	-	-	-
A2	-	-	-	-	-	-
A3	-	-	WT	-	-	-
A4	-	-	-	-	-	-
A5	-	-	-	-	-	-
A6	-	-	-	-	-	-
A7	-	-	-	-	-	-
A8	-	-	-	-	-	-
A9	-	-	-	-	-	-
A11	-	-	Q8	-	-	-
A12	-	-	-	-	-	-
A13	-	-	-	-	-	-
A14	63P	65R 70R 108I 181C 184V 219E 333E 211K 214F	F46 I90 M90	R70a I108 C181 K211 E219	V82b-96 I82-4	R70a I108-100 M178-4 C181-100 K211-100 E219-100
A15	-	-	V82b I82	-	V82b-62 I82-38	-

			T82 V84		I54-95	
A16	-	-	-	-	-	-
A17	-	-	-	-	-	-
A18	-	-	-	-	-	-
A19	-	-	I36	-	I36-5	-
A20	-	-	WT	-		-
A21	-	-	-	-	-	-
A23	-	-	-	-	-	-
A24	-	-	-	-	-	-
A25	-	-	-	-	-	-
A26	-	-	-	-	-	-
A27	36I	41L 68G 74V 151M 184V 190S 211K 215F	-	M151 A190 K211 F215	-	M151-100 A190-21 K211-100 F215-74
A28			-	-	-	-
A31			WT	-	-	-
A32	20R 36I 63P	67N 116Y 151M 190A 219E	I36	M151 A190 E219 T238	I46-5 A71-100	M151-94 A190-100 E219-22 T238-100

A33			V82b T82 A82	-	V48-5 A82-100	-
A34			N30	-	N30-100	-
A37	-	-	M46b I46 V82b A82	-	I46-100 A82-100	-
A38	-	-	V54 F82	-	R8-95 Q8-5	-
A39	63P 71V V77I 93L	65R 115F 184V	F82 M90	WT	V71-100	-
A40	-	-	I82	-	I82-70	-
A41	-	-	I46	-	D30-95 N30-5 I46-90	-
A44	-	-	-	-	-	-
A45	-	-	-	-	-	-
A46	-	-	-	-	-	-
A47	-	-	-	-	-	-
A48	-	-	-	-	-	-
A49	-	-	-	-	-	-
A51	-	-	-	-	-	-
A52	-	-	V82b A82	-	A82-100 M90-100	-

			L90a M90			
A53	-	-	V54 V82b A82 M90	-	V54-95 A82-100 M90-100	-
A54	-	-	-	-	-	-
A55	-	-	-	-	-	-
A56	-	-	-	-	-	-
A57	-	-		-	-	-
A58	-	-	I84 V84 M90	-	V84-100 M90-100	-
A59	-	-	I36 A82	-	I36-100 A82-100	-
A60				D69 E101 A190 K211		D69-100 E101-100 A190-100 K211-100
A61	L10I M46I I54V V77I V82A/F	M41L E44D A62V 69INS V118I L210W T215Y K103N	Q8	V62 Ins69j C181 S190 K211		V62-100 Ins69j-100 C181-76 S190-24 K211-85

		Y181C G190A/S				
A62						
A63						
A64						
A65						

Table 1.- Comparison of genotypes obtained from hospital analyse, microarray hybridization and molecular cloning of HIV-PR and -RT from infected patients.

Signals	**Classification accuracy (% of signals) before filtering**		
	Training Set	**Test Set**	
		PR	**RT**
Correct	93.04	84.42	85.27
False Positives (FP)	5.09	9.06	9.80
False Negatives (FN)	1.33	5.41	2.72
Undefined (UD)	0.54	1.11	2.15

Table 2.- Classification accuracy (before filtering) of the microarrays from the training and test sets, by comparing the experimental hybridization signals with those derived from the theoretical hybridization tables.

	Discarded features (%)	
Quality control step	**Training set**	**Test set**
1: Spot filter	0.20 (spots)	1.27% (RT spots) to 4.05% (PR spots)
2: Probe filter 1 (overlapped)	9.09 (probes overlapping >25%)	9.09 (probes overlapping >25%)
3: Probe filter 2 (duplications)	10.13 (duplicated probes)	10.13 (duplicated probes)
4: Array filter	4.46 (arrays)	0% (RT arrays) to 4.16% (PR arrays)

Table 3.- Quality control summary.

Signals	**Step 1: Spot filter**			**Step 2: Probe filter 1 (overlapped)**			**Step 3: Probe filter 2 (duplications)**		
	TS	**SS**		**TS**	**SS**		**TS**	**SS**	
		PR	**RT**		**PR**	**RT**		**PR**	**RT**
Correct	94.01	86.00	86.09	96.37	92.42	88.48	95.43	92.14	89.24
FP	3.99	7.62	9.54	2.63	4.70	7.14	3.06	4.89	6.87
FN	1.61	5.39	2.50	0.52	2.21	2.40	1.02	2.13	2.26
UD	0.40	0.99	1.87	0.48	0.67	1.99	0.49	0.85	1.64

Table 4.-Classification accuracy including 4-steps filtering. TS training set, SS test set. PR protease, RT Retrotranscriptase (continue in next page*).

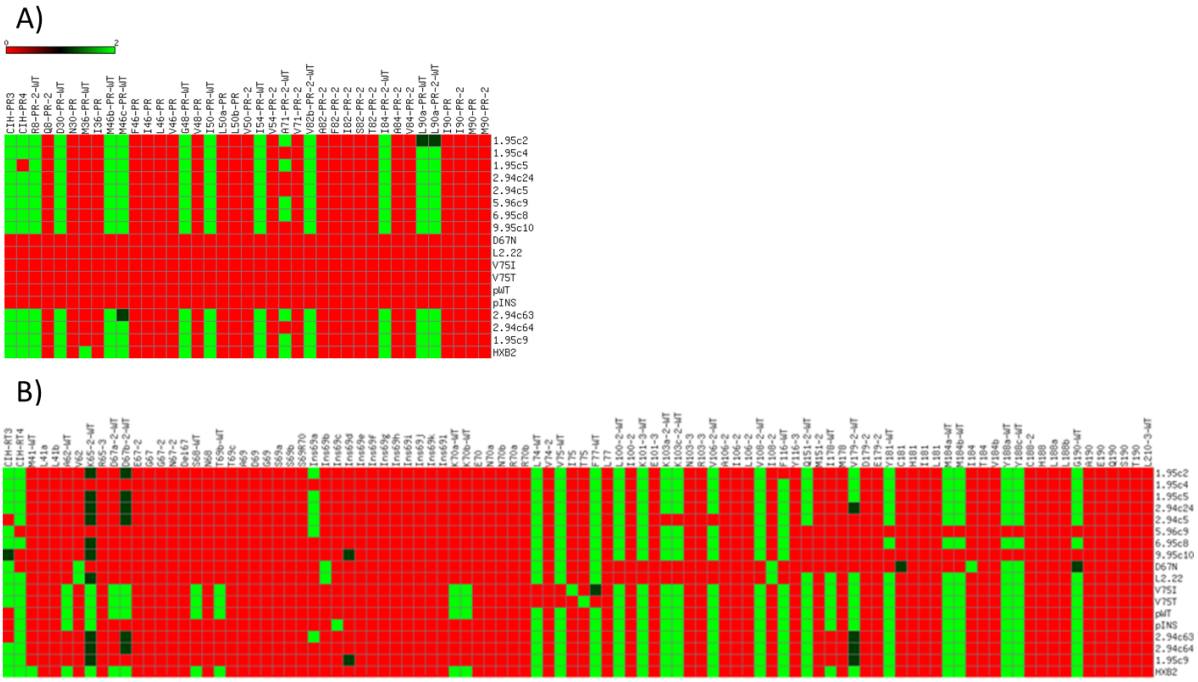
Signals	Step 4: Array filter		
	TS	SS	
		PR	RT
Correct	**96.33**	**93.53**	**89.24**
FP	2.19	3.84	6.87
FN	1.07	2.16	2.26
UD	0.41	0.48	1.64

Table 4'.- Classification accuracy including 4-steps filtering. TS training set, SS test set. PR protease, RT Retrotranscriptase.

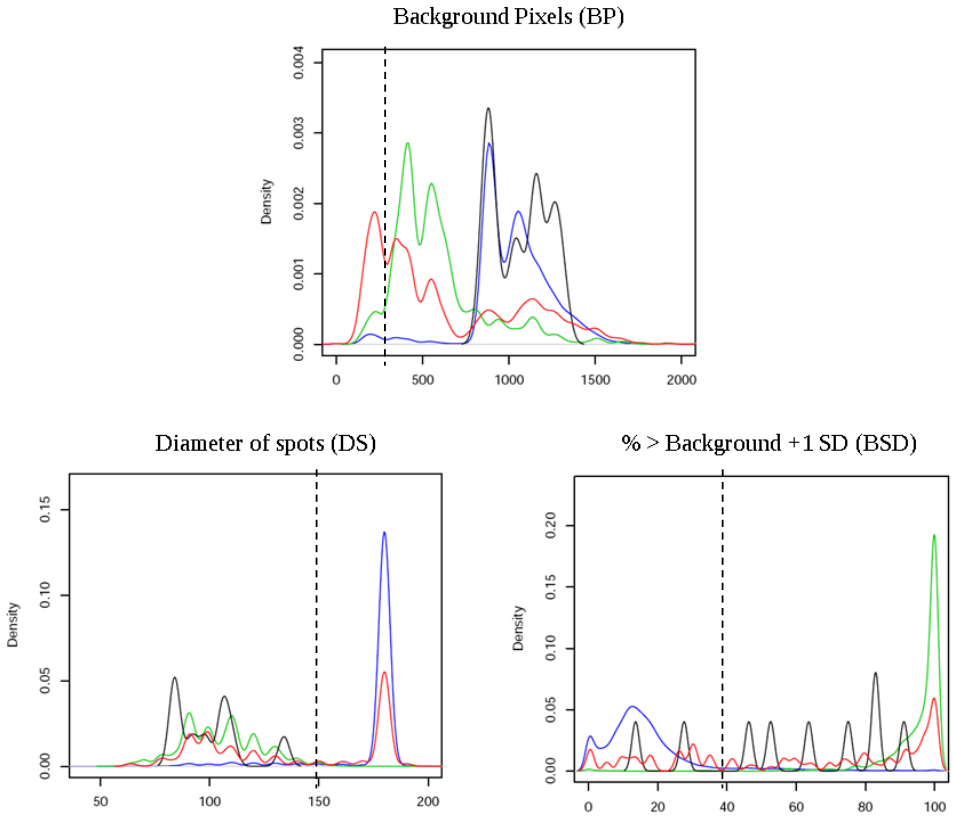
Range of minority variants within the quasispecies (%)	Fraction of positions (PR + RT)
1.00 – 4.99	15/970 (1.55%)
5.00 – 9.99	34/970 (3.50%)
10.00 – 19.99	24/970 (2.47%)

Table 5.- Minority subpopulations present in the clinical samples derived from clonal analysis.

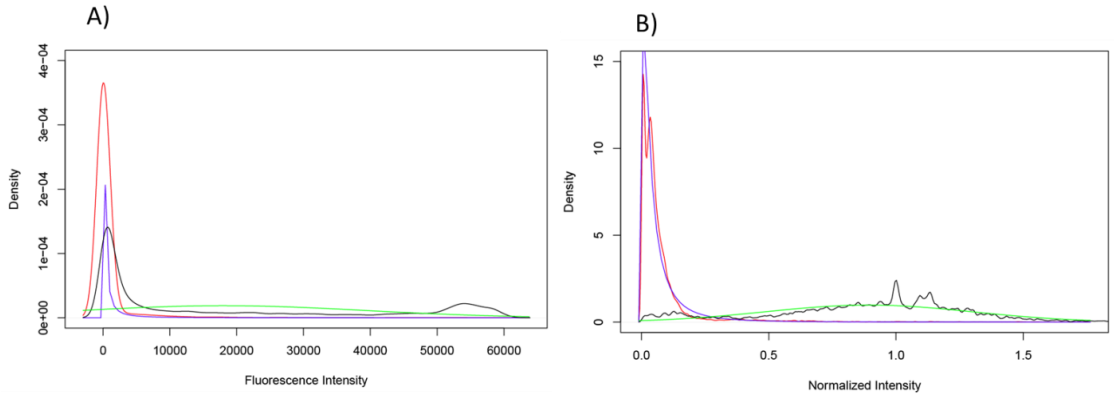
Supplementary Material



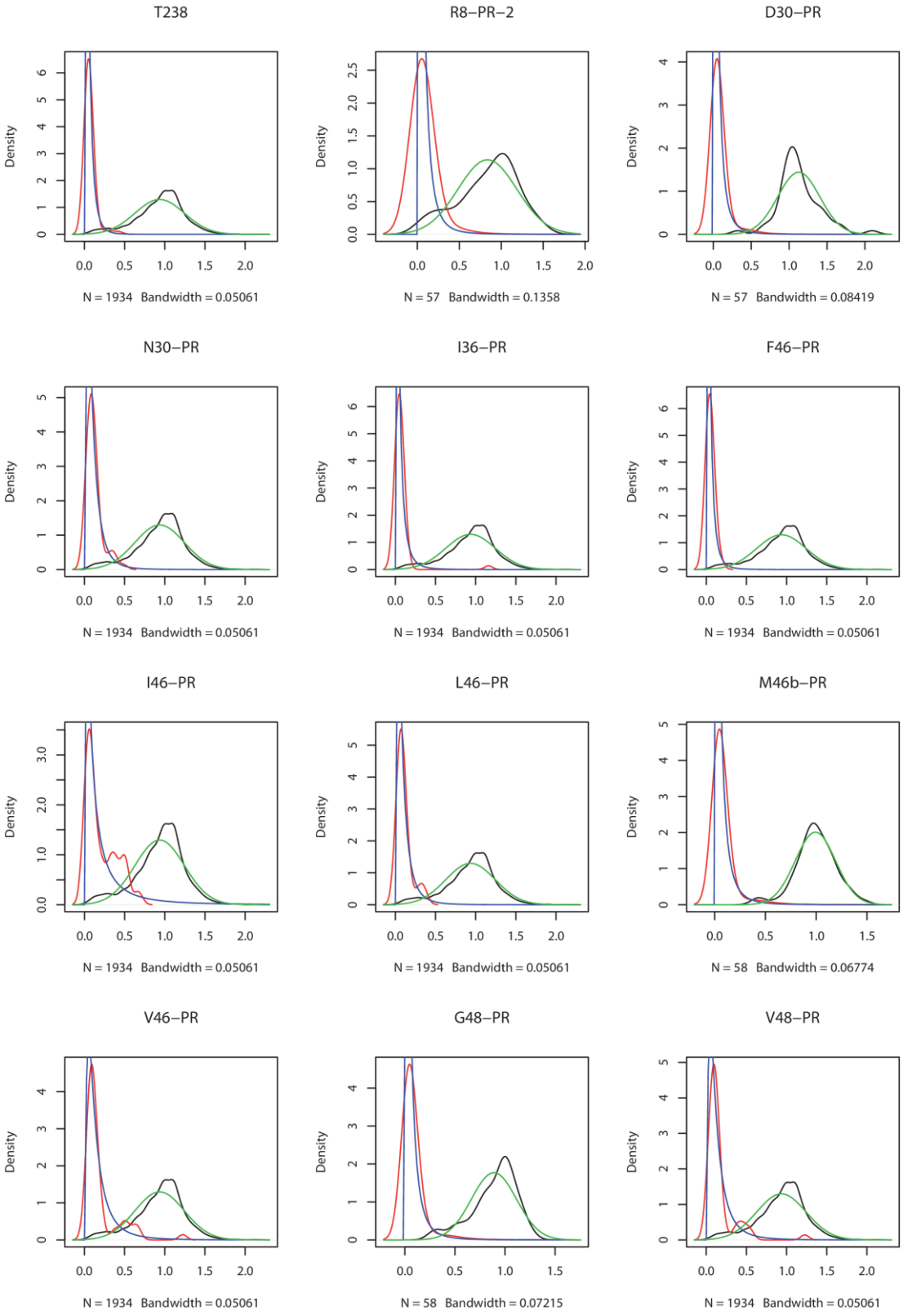
Supplementary Fig. S1.- Theoretical hybridization tables of pure clonal samples belonging to the training set, without filtering. A) Columns, probes belonging to the PR region; rows, samples; **B)** Columns, probes belonging to the RT region; rows, samples. Legend: green, expected hybridization; red, not expected hybridization; black: partial hybridization (one mismatch between probe and target is allowed at the 5' or 3' position of the hybridizing sequence).

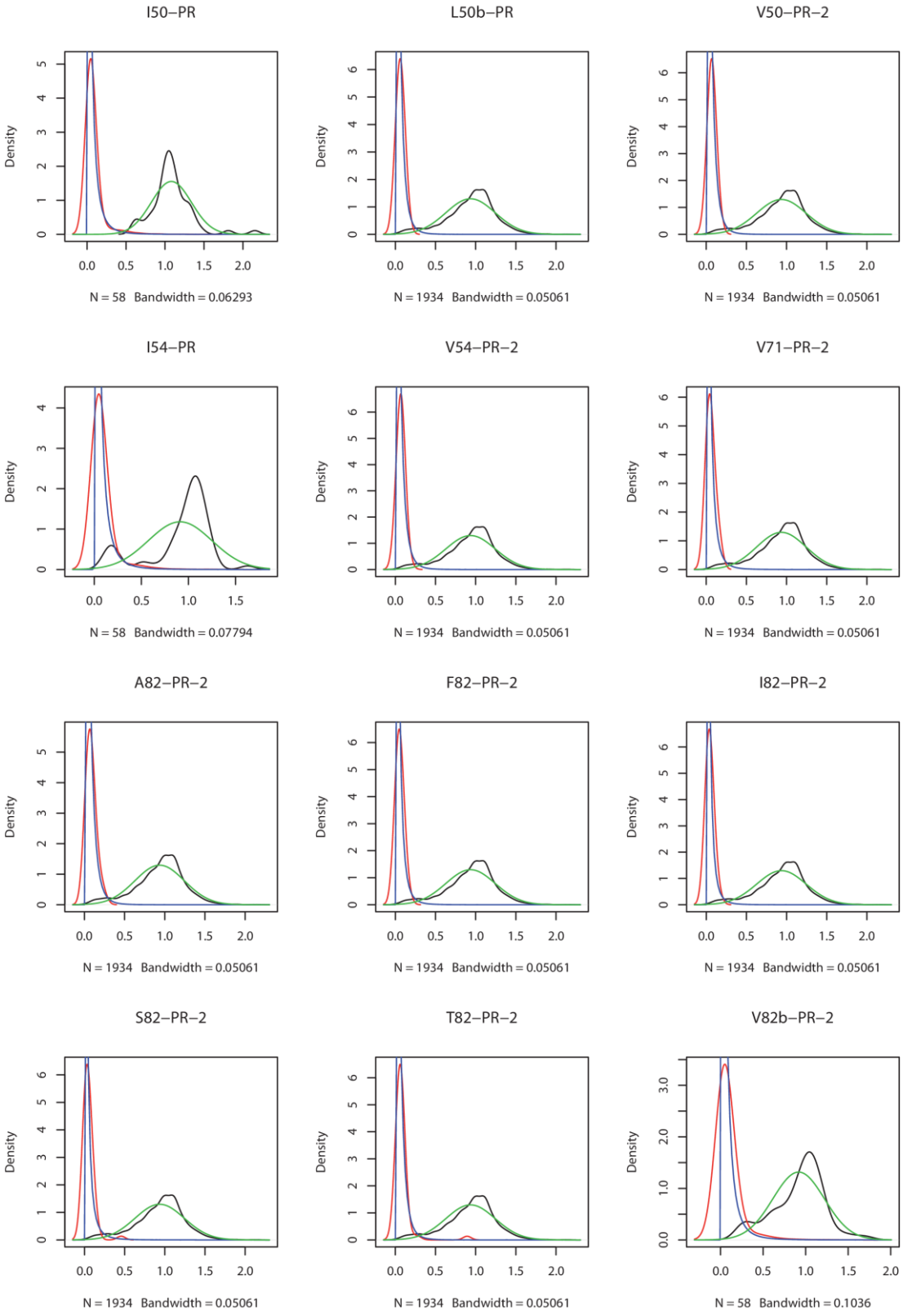


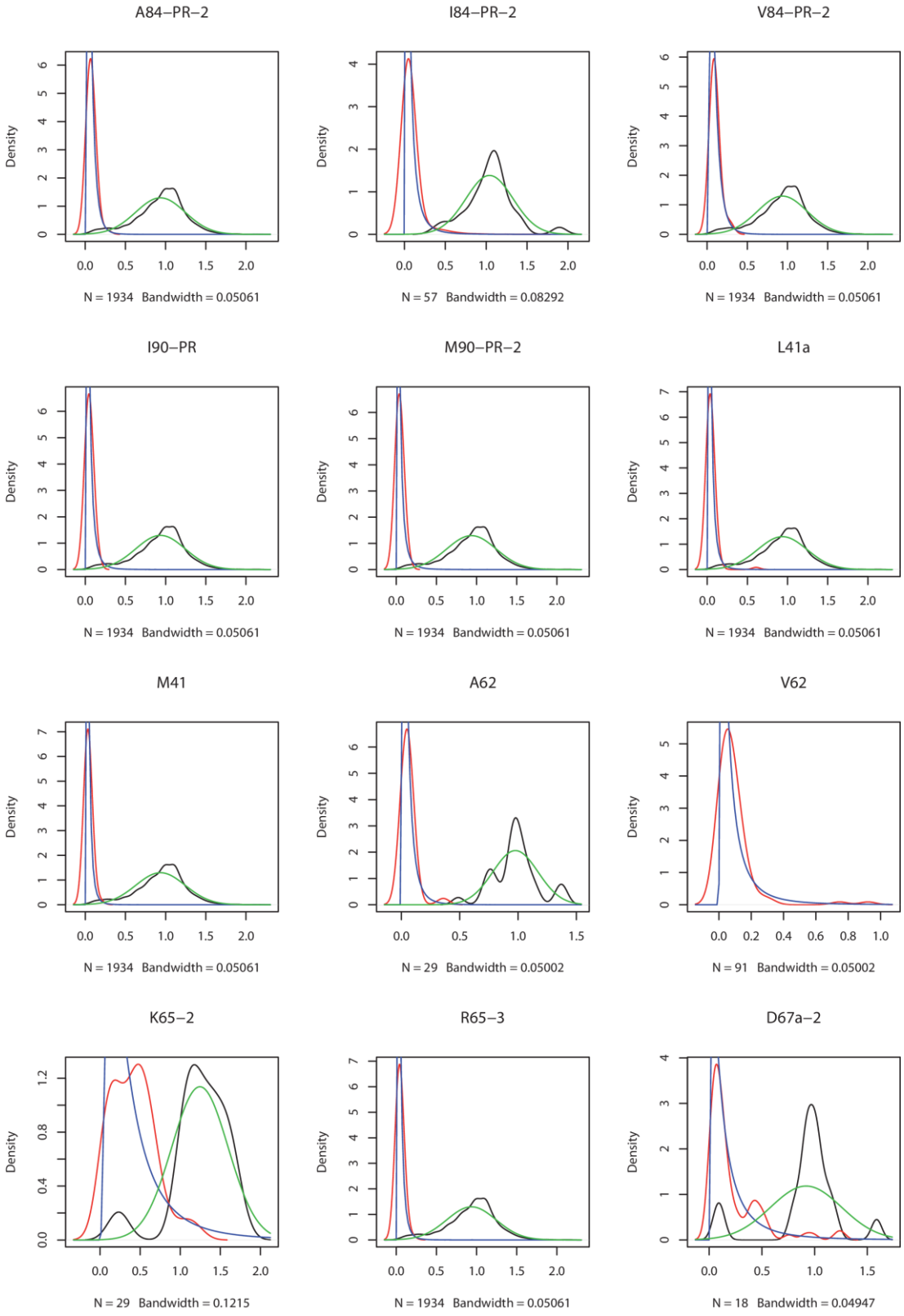
Supplementary Fig. S2.- Density of the raw data for three selected variables included in the microarray lecture obtained by GenePix and ScanArray scanners. Legend: green, true positive (TP) signal; red, false positive (FP); blue, true negative (TN); black, false negative (FN).

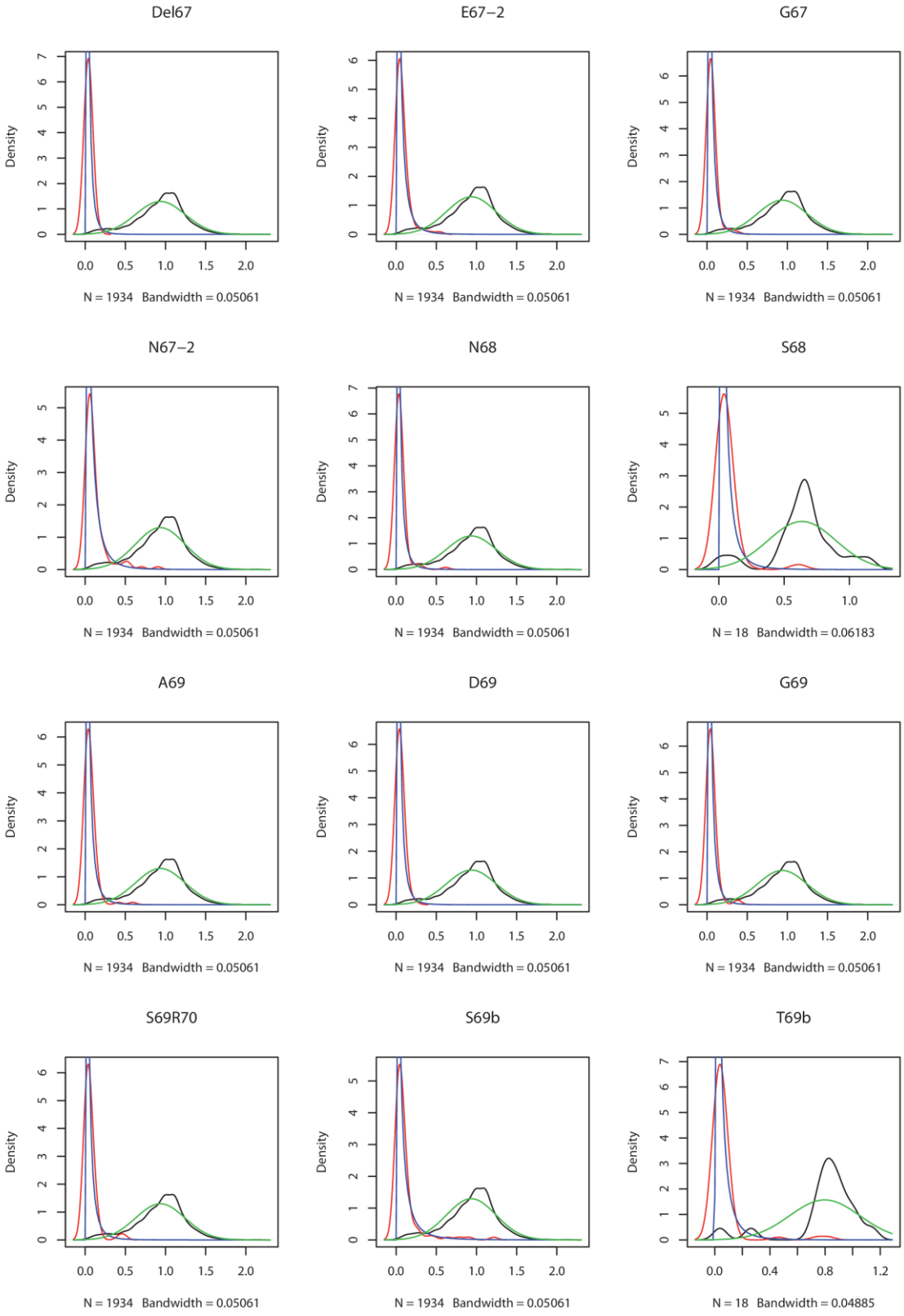


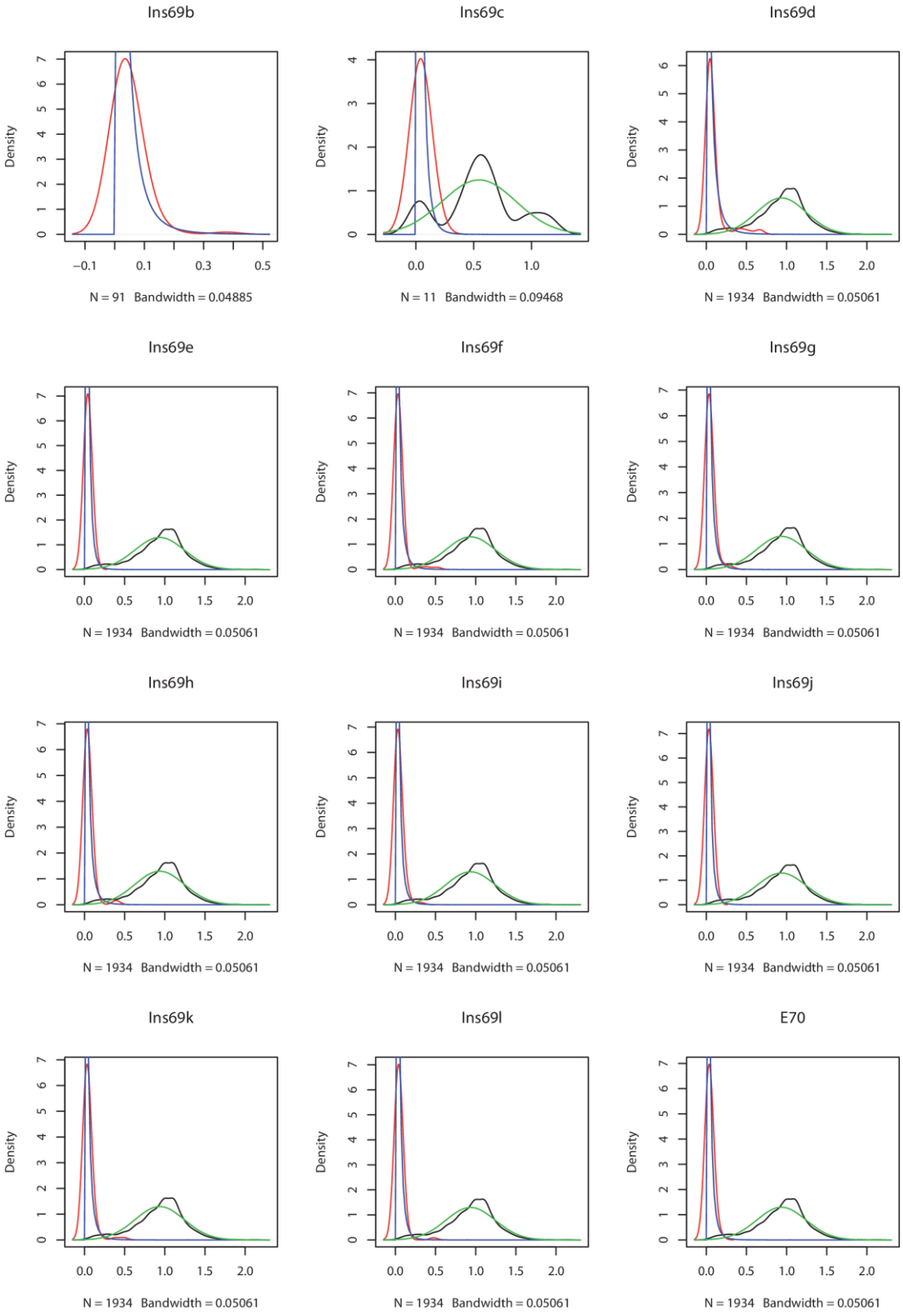
Supplementary Fig. S3.- Test of the normalization performance of data from the training set. A) Density of positive and negative raw data. **B)** Normalized data using the mean positive signal/array as the selected factor. Distributions of positive signals are centred in value 1 of intensity, and distributions of negative signals close to zero as expected. Legend: black, positive signal; red, negative signal; dotted black, positive fit; dotted red, *negative fit.*

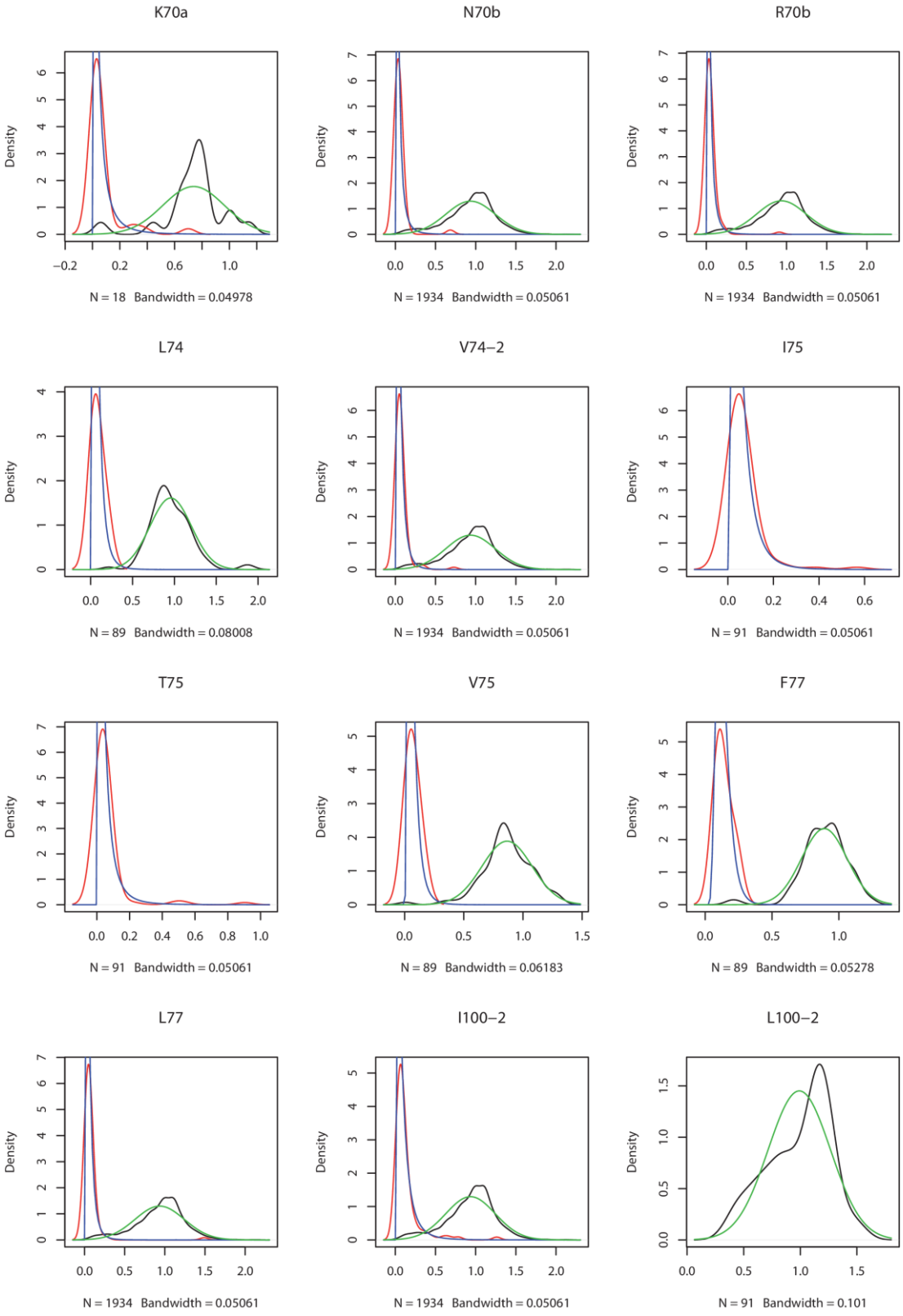

T238
Density
N = 1934 Bandwidth = 0.05061
R8–PR–2
Density
N = 57 Bandwidth = 0.1358
D30–PR
Density
N = 57 Bandwidth = 0.08419
N30–PR
Density
N = 1934 Bandwidth = 0.05061
I36–PR
Density
N = 1934 Bandwidth = 0.05061
F46–PR
Density
N = 1934 Bandwidth = 0.05061
I46–PR
Density
N = 1934 Bandwidth = 0.05061
L46–PR
Density
N = 1934 Bandwidth = 0.05061
M46b–PR
Density
N = 58 Bandwidth = 0.06774
V46–PR
Density
N = 1934 Bandwidth = 0.05061
G48–PR
Density
N = 58 Bandwidth = 0.07215
V48–PR
Density
N = 1934 Bandwidth = 0.05061

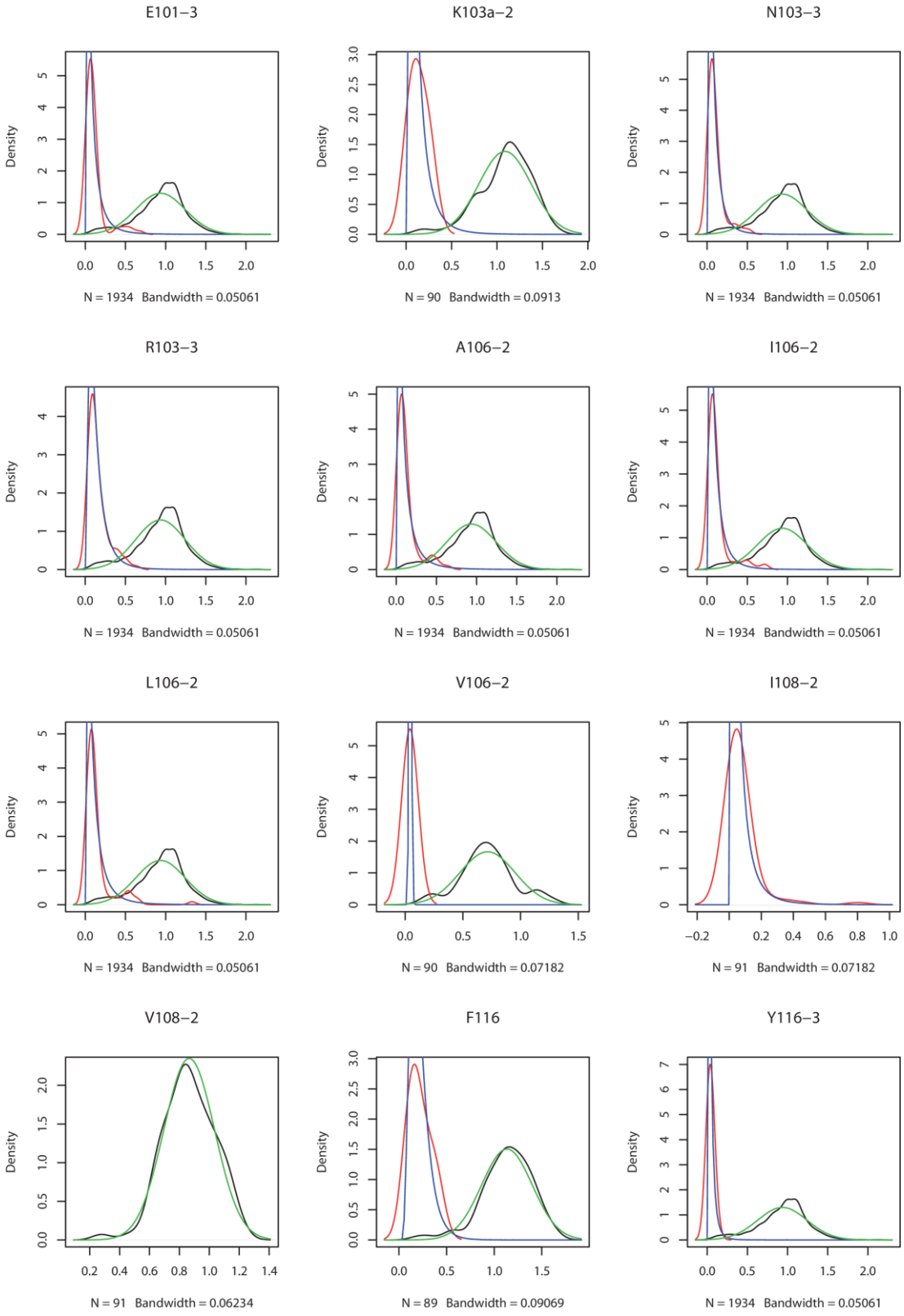

I50–PR
Density
N = 58 Bandwidth = 0.06293
L50b–PR
Density
N = 1934 Bandwidth = 0.05061
V50–PR–2
Density
N = 1934 Bandwidth = 0.05061
I54–PR
Density
N = 58 Bandwidth = 0.07794
V54–PR–2
Density
N = 1934 Bandwidth = 0.05061
V71–PR–2
Density
N = 1934 Bandwidth = 0.05061
A82–PR–2
Density
N = 1934 Bandwidth = 0.05061
F82–PR–2
Density
N = 1934 Bandwidth = 0.05061
I82–PR–2
Density
N = 1934 Bandwidth = 0.05061
S82–PR–2
Density
N = 1934 Bandwidth = 0.05061
T82–PR–2
Density
N = 1934 Bandwidth = 0.05061
V82b–PR–2
Density
N = 58 Bandwidth = 0.1036

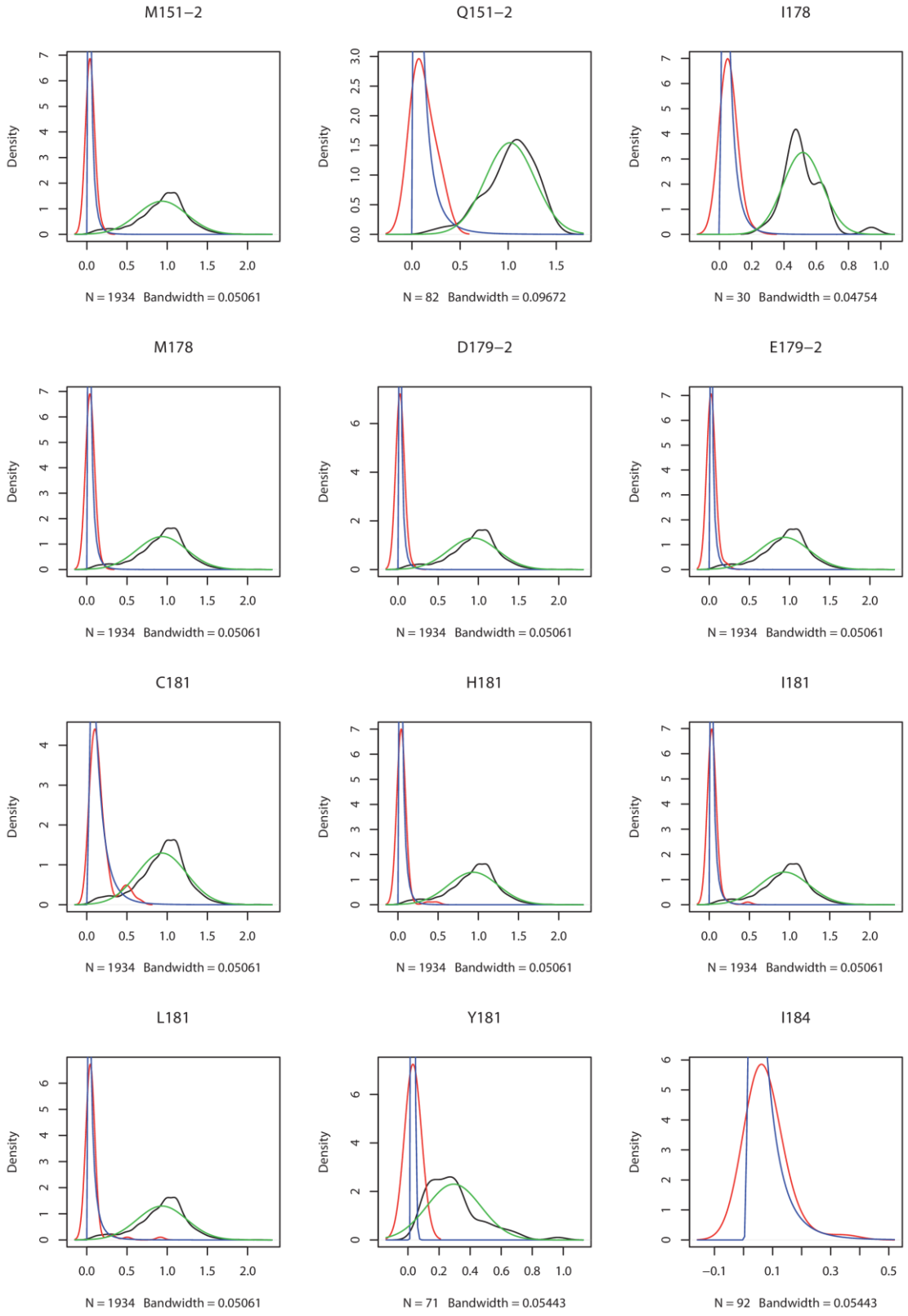

A84–PR–2
Density
N = 1934 Bandwidth = 0.05061
I84–PR–2
Density
N = 57 Bandwidth = 0.08292
V84–PR–2
Density
N = 1934 Bandwidth = 0.05061
I90–PR
Density
N = 1934 Bandwidth = 0.05061
M90–PR–2
Density
N = 1934 Bandwidth = 0.05061
L41a
Density
N = 1934 Bandwidth = 0.05061
M41
Density
N = 1934 Bandwidth = 0.05061
A62
Density
N = 29 Bandwidth = 0.05002
V62
Density
N = 91 Bandwidth = 0.05002
K65–2
Density
N = 29 Bandwidth = 0.1215
R65–3
Density
N = 1934 Bandwidth = 0.05061
D67a–2
Density
N = 18 Bandwidth = 0.04947

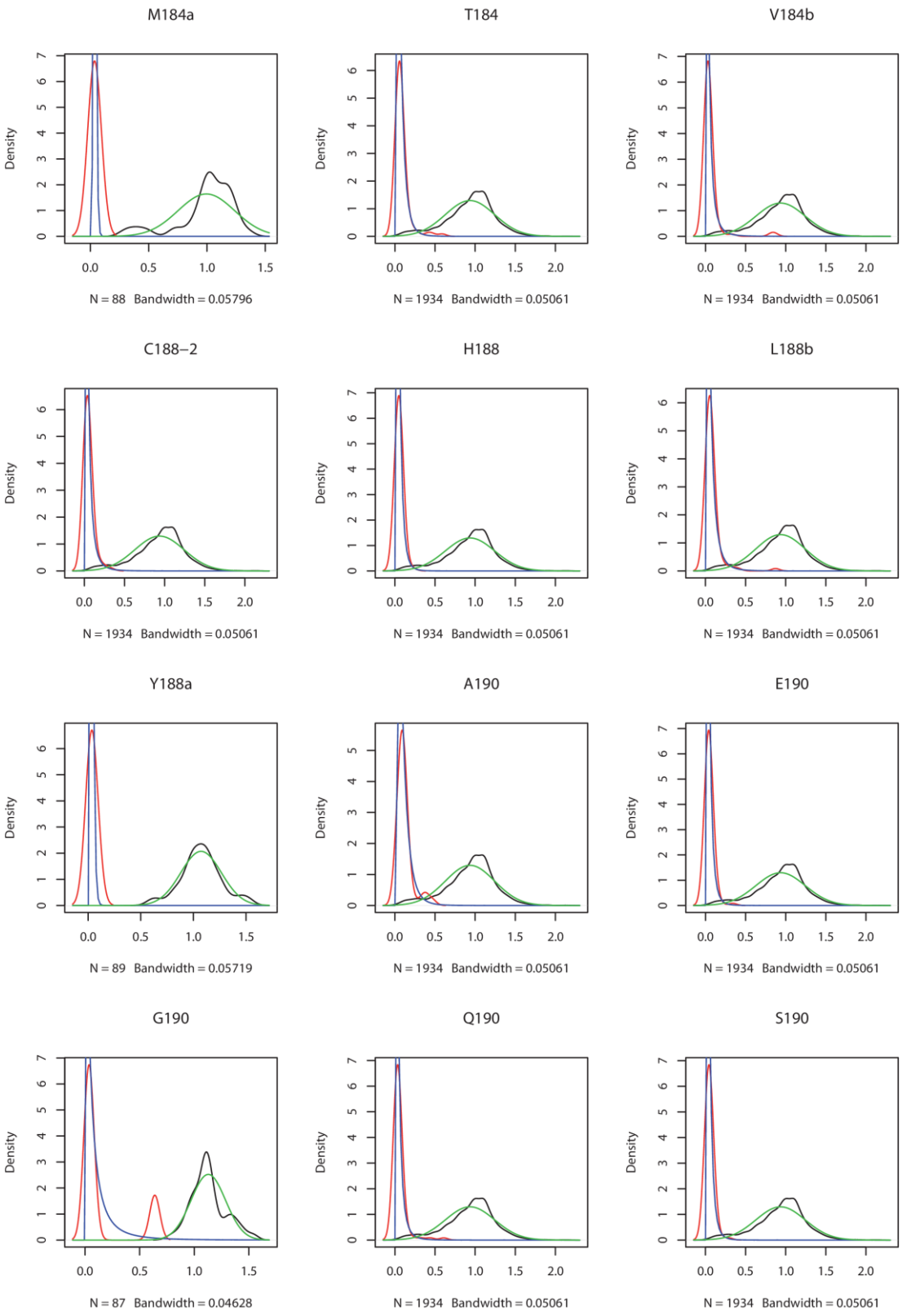

Del67
E67−2
G67
N67−2
N68
S68
A69
D69
G69
S69R70
S69b
T69b
Density
N = 1934 Bandwidth = 0.05061
N = 18 Bandwidth = 0.06183
N = 18 Bandwidth = 0.04885

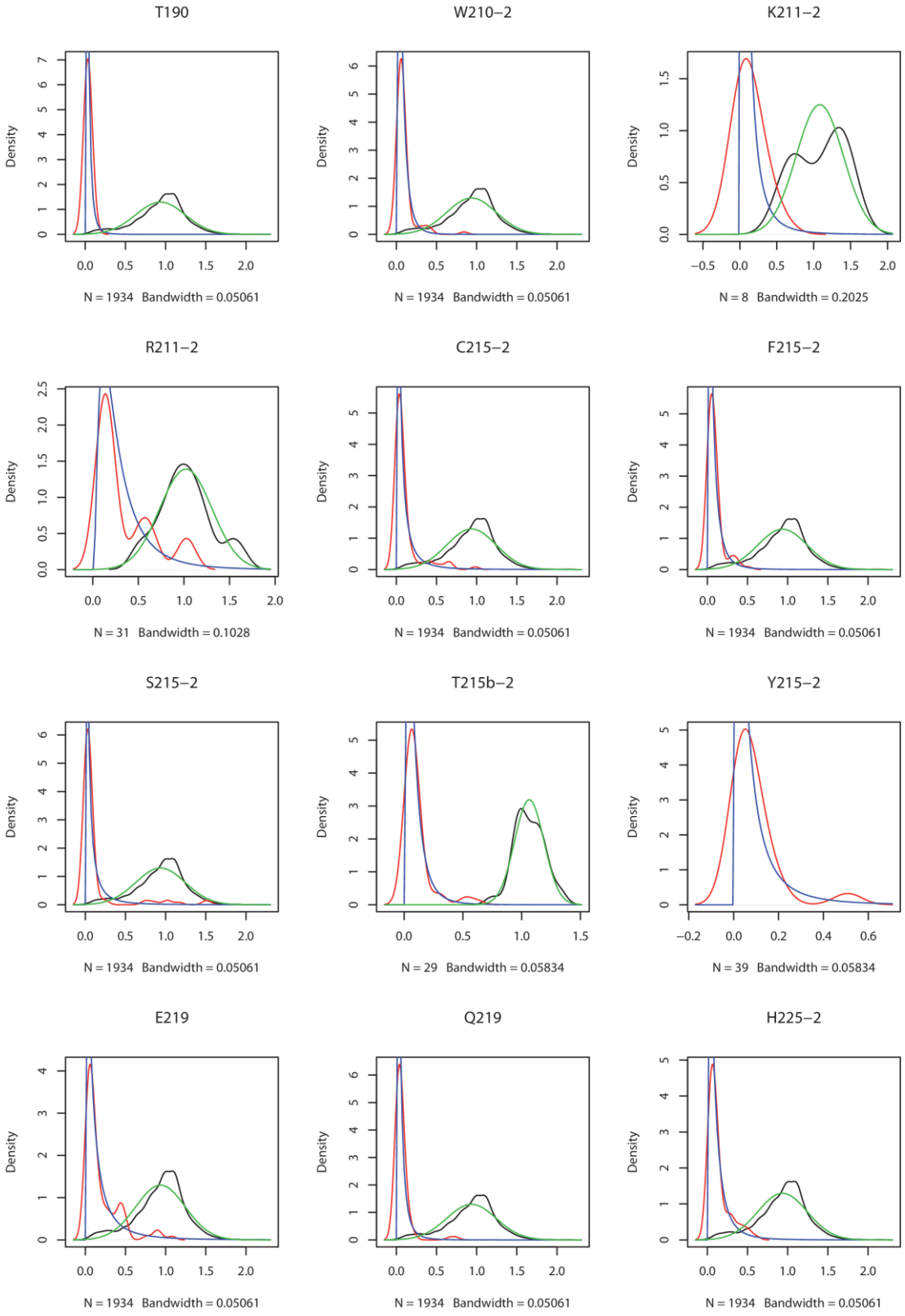

Ins69b
Density
N = 91 Bandwidth = 0.04885
Ins69c
Density
N = 11 Bandwidth = 0.09468
Ins69d
Density
N = 1934 Bandwidth = 0.05061
Ins69e
Density
N = 1934 Bandwidth = 0.05061
Ins69f
Density
N = 1934 Bandwidth = 0.05061
Ins69g
Density
N = 1934 Bandwidth = 0.05061
Ins69h
Density
N = 1934 Bandwidth = 0.05061
Ins69i
Density
N = 1934 Bandwidth = 0.05061
Ins69j
Density
N = 1934 Bandwidth = 0.05061
Ins69k
Density
N = 1934 Bandwidth = 0.05061
Ins69l
Density
N = 1934 Bandwidth = 0.05061
E70
Density
N = 1934 Bandwidth = 0.05061

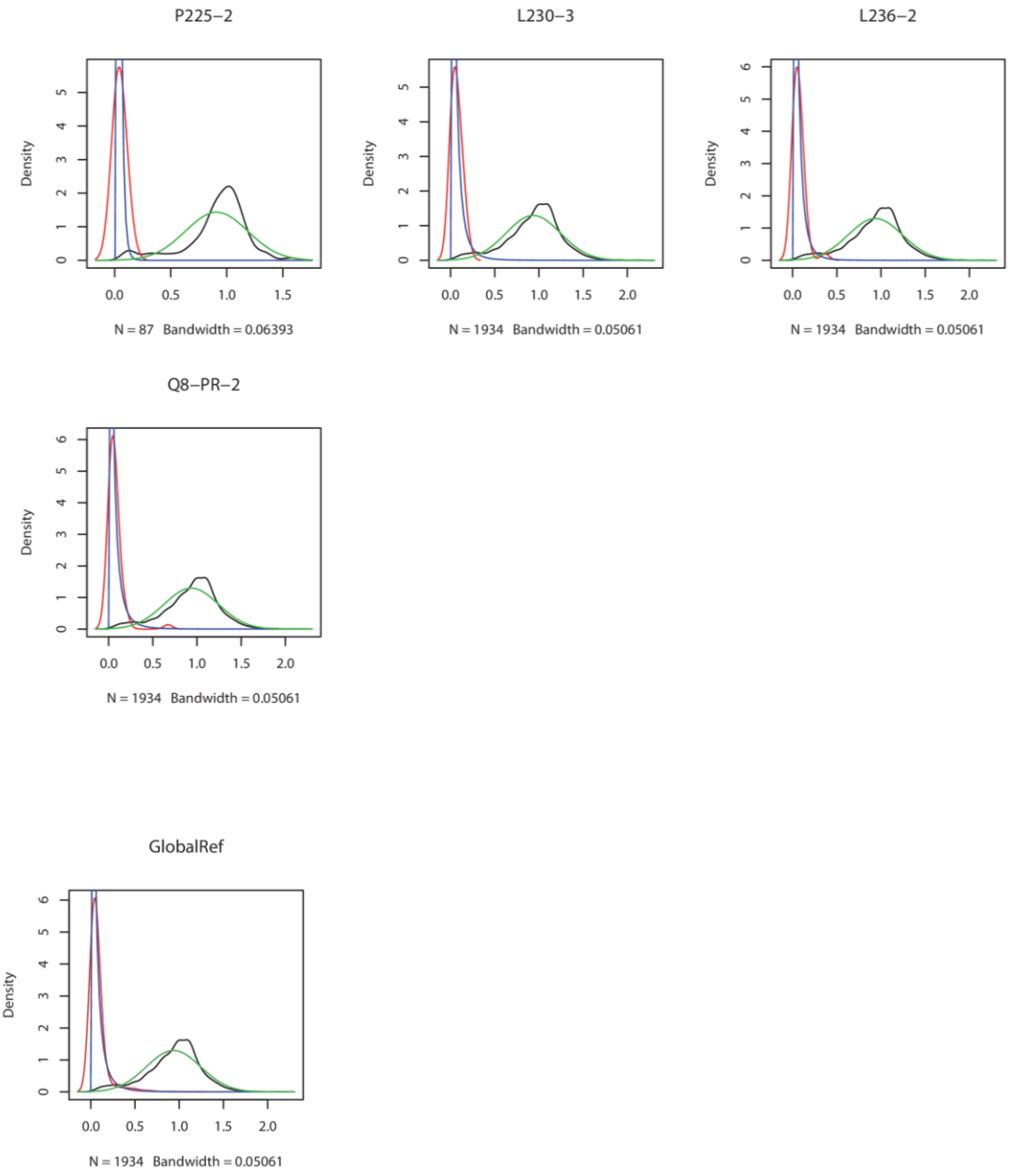

K70a
Density
N = 18 Bandwidth = 0.04978
N70b
Density
N = 1934 Bandwidth = 0.05061
R70b
Density
N = 1934 Bandwidth = 0.05061
L74
Density
N = 89 Bandwidth = 0.08008
V74–2
Density
N = 1934 Bandwidth = 0.05061
I75
Density
N = 91 Bandwidth = 0.05061
T75
Density
N = 91 Bandwidth = 0.05061
V75
Density
N = 89 Bandwidth = 0.06183
F77
Density
N = 89 Bandwidth = 0.05278
L77
Density
N = 1934 Bandwidth = 0.05061
I100–2
Density
N = 1934 Bandwidth = 0.05061
L100–2
Density
N = 91 Bandwidth = 0.101


E101–3
Density
N = 1934 Bandwidth = 0.05061
K103a–2
Density
N = 90 Bandwidth = 0.0913
N103–3
Density
N = 1934 Bandwidth = 0.05061
R103–3
Density
N = 1934 Bandwidth = 0.05061
A106–2
Density
N = 1934 Bandwidth = 0.05061
I106–2
Density
N = 1934 Bandwidth = 0.05061
L106–2
Density
N = 1934 Bandwidth = 0.05061
V106–2
Density
N = 90 Bandwidth = 0.07182
I108–2
Density
N = 91 Bandwidth = 0.07182
V108–2
Density
N = 91 Bandwidth = 0.06234
F116
Density
N = 89 Bandwidth = 0.09069
Y116–3
Density
N = 1934 Bandwidth = 0.05061

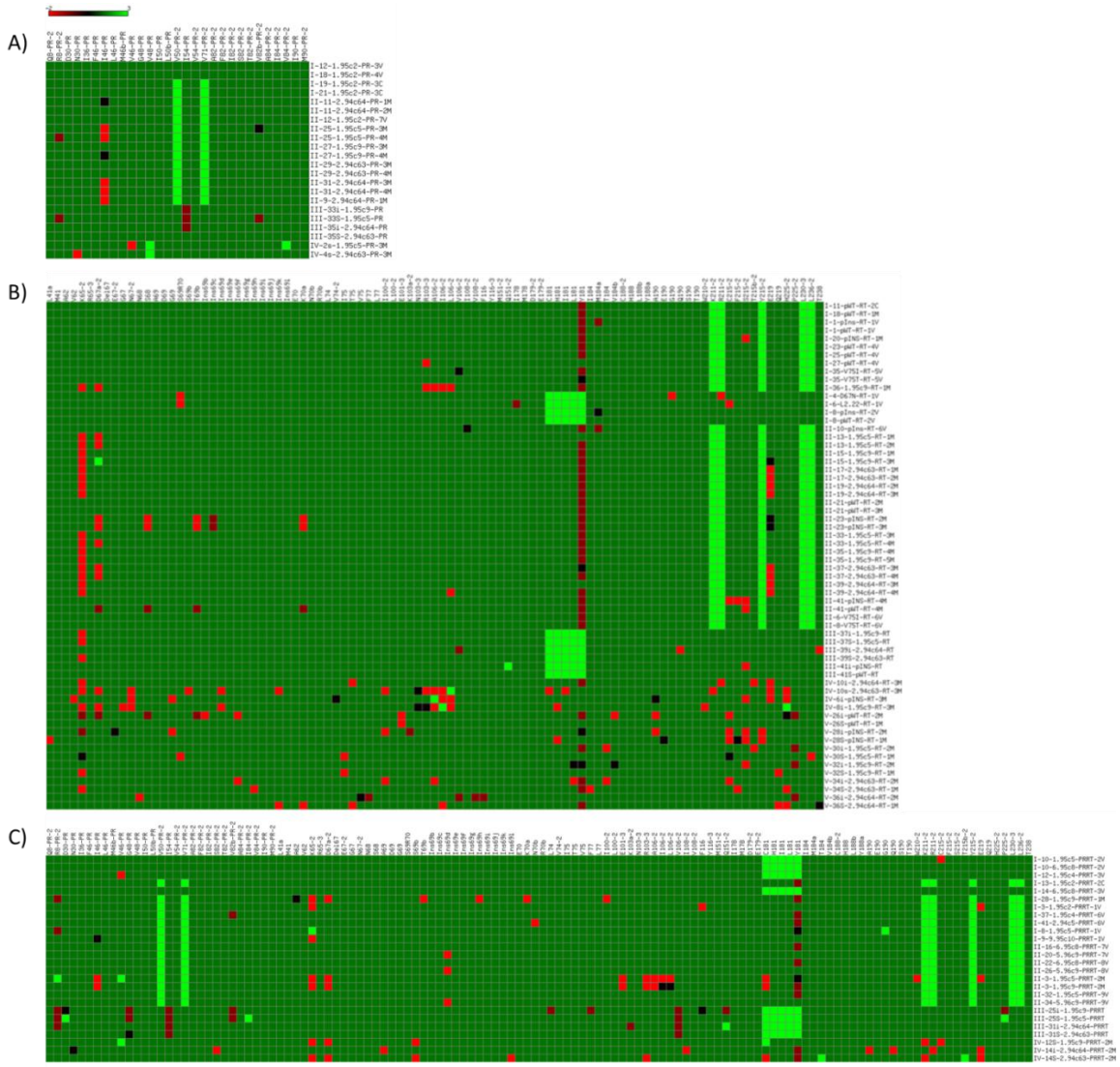

M151–2
Density
N = 1934 Bandwidth = 0.05061
Q151–2
Density
N = 82 Bandwidth = 0.09672
I178
Density
N = 30 Bandwidth = 0.04754
M178
Density
N = 1934 Bandwidth = 0.05061
D179–2
Density
N = 1934 Bandwidth = 0.05061
E179–2
Density
N = 1934 Bandwidth = 0.05061
C181
Density
N = 1934 Bandwidth = 0.05061
H181
Density
N = 1934 Bandwidth = 0.05061
I181
Density
N = 1934 Bandwidth = 0.05061
L181
Density
N = 1934 Bandwidth = 0.05061
Y181
Density
N = 71 Bandwidth = 0.05443
I184
Density
N = 92 Bandwidth = 0.05443


M184a
Density
N = 88 Bandwidth = 0.05796
T184
Density
N = 1934 Bandwidth = 0.05061
V184b
Density
N = 1934 Bandwidth = 0.05061
C188–2
Density
N = 1934 Bandwidth = 0.05061
H188
Density
N = 1934 Bandwidth = 0.05061
L188b
Density
N = 1934 Bandwidth = 0.05061
Y188a
Density
N = 89 Bandwidth = 0.05719
A190
Density
N = 1934 Bandwidth = 0.05061
E190
Density
N = 1934 Bandwidth = 0.05061
G190
Density
N = 87 Bandwidth = 0.04628
Q190
Density
N = 1934 Bandwidth = 0.05061
S190
Density
N = 1934 Bandwidth = 0.05061

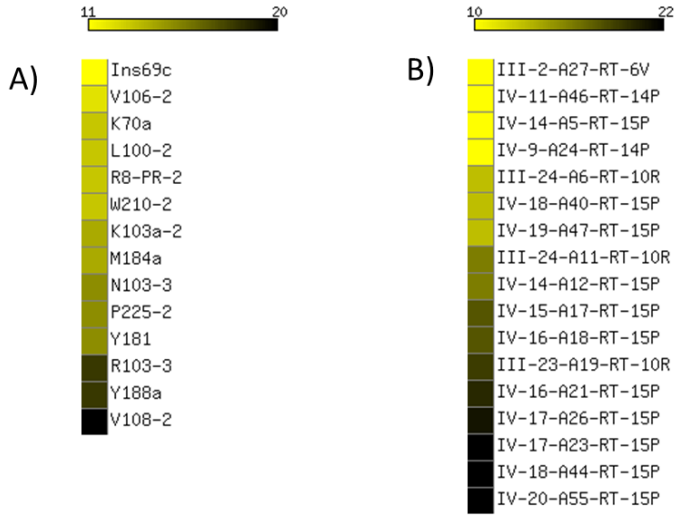

T190
Density
N = 1934 Bandwidth = 0.05061
W210–2
Density
N = 1934 Bandwidth = 0.05061
K211–2
Density
N = 8 Bandwidth = 0.2025
R211–2
Density
N = 31 Bandwidth = 0.1028
C215–2
Density
N = 1934 Bandwidth = 0.05061
F215–2
Density
N = 1934 Bandwidth = 0.05061
S215–2
Density
N = 1934 Bandwidth = 0.05061
T215b–2
Density
N = 29 Bandwidth = 0.05834
Y215–2
Density
N = 39 Bandwidth = 0.05834
E219
Density
N = 1934 Bandwidth = 0.05061
Q219
Density
N = 1934 Bandwidth = 0.05061
H225–2
Density
N = 1934 Bandwidth = 0.05061



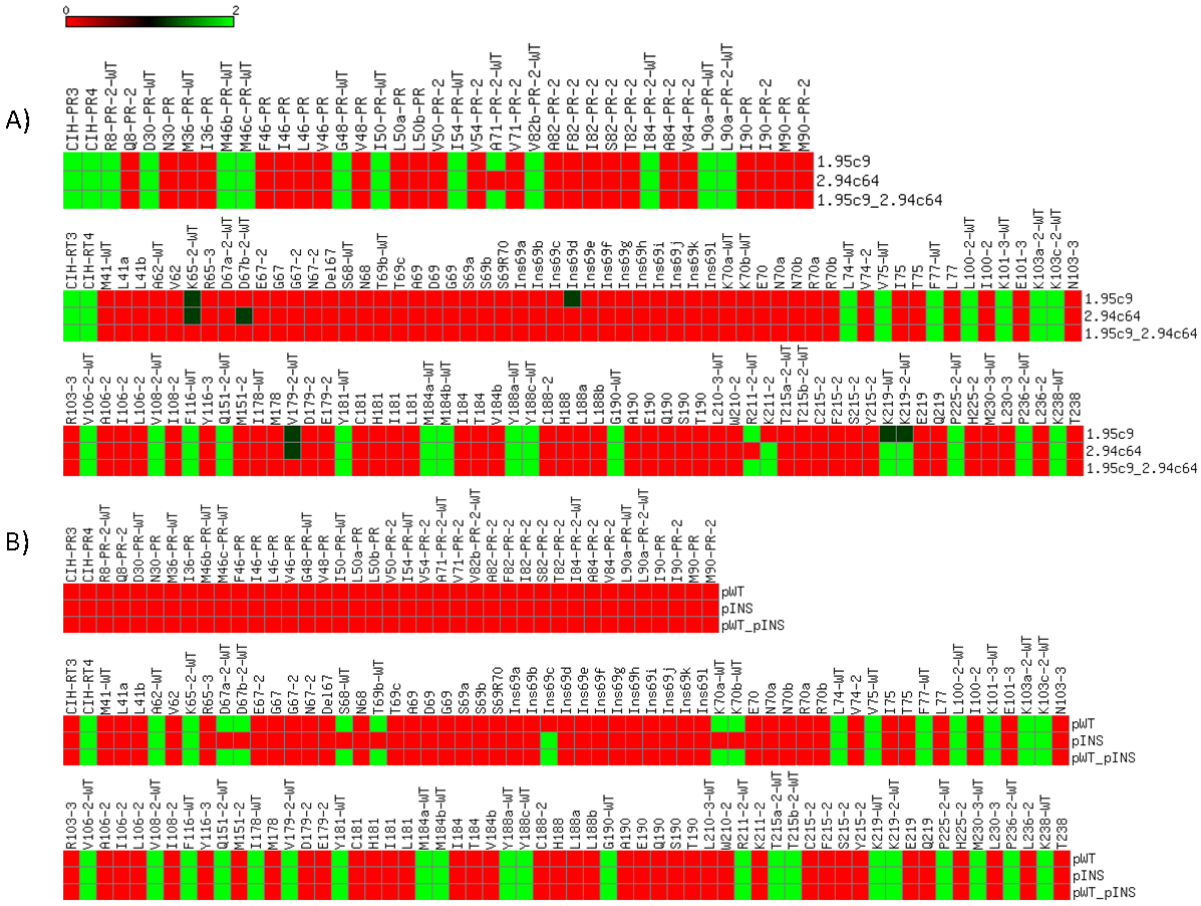
Supplementary Fig. S4.- Function distributions of positive and negative signals (signals/probe and global references)



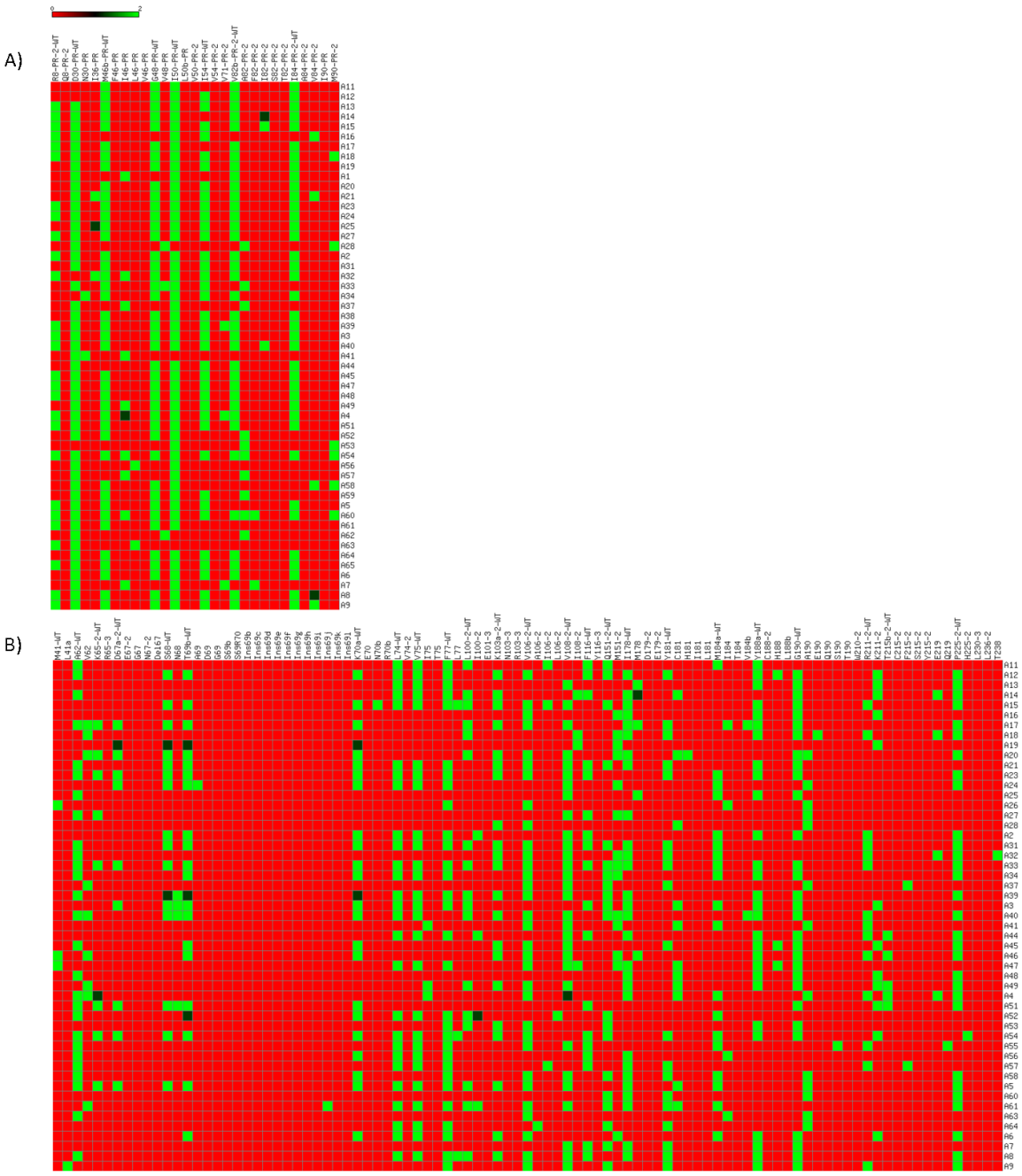
Supplementary Fig. S5.- Classification accuracy in the pure training set. Columns, probes included in Virochip 4.0 that passed the four filtering steps; rows, samples. **A)** Hybridization of the PR region of the samples; **B)** Hybridization of the RT; **C)** Hybridization of amplicons including the PR and RT. Legend: dark green, correctly classified signal (TP or TN); dark red, FN signal; red, FP signal; black, UD spot; light green, no data (due to spots discarded during the quality control or to the absence of certain probes in different versions of the microarray).



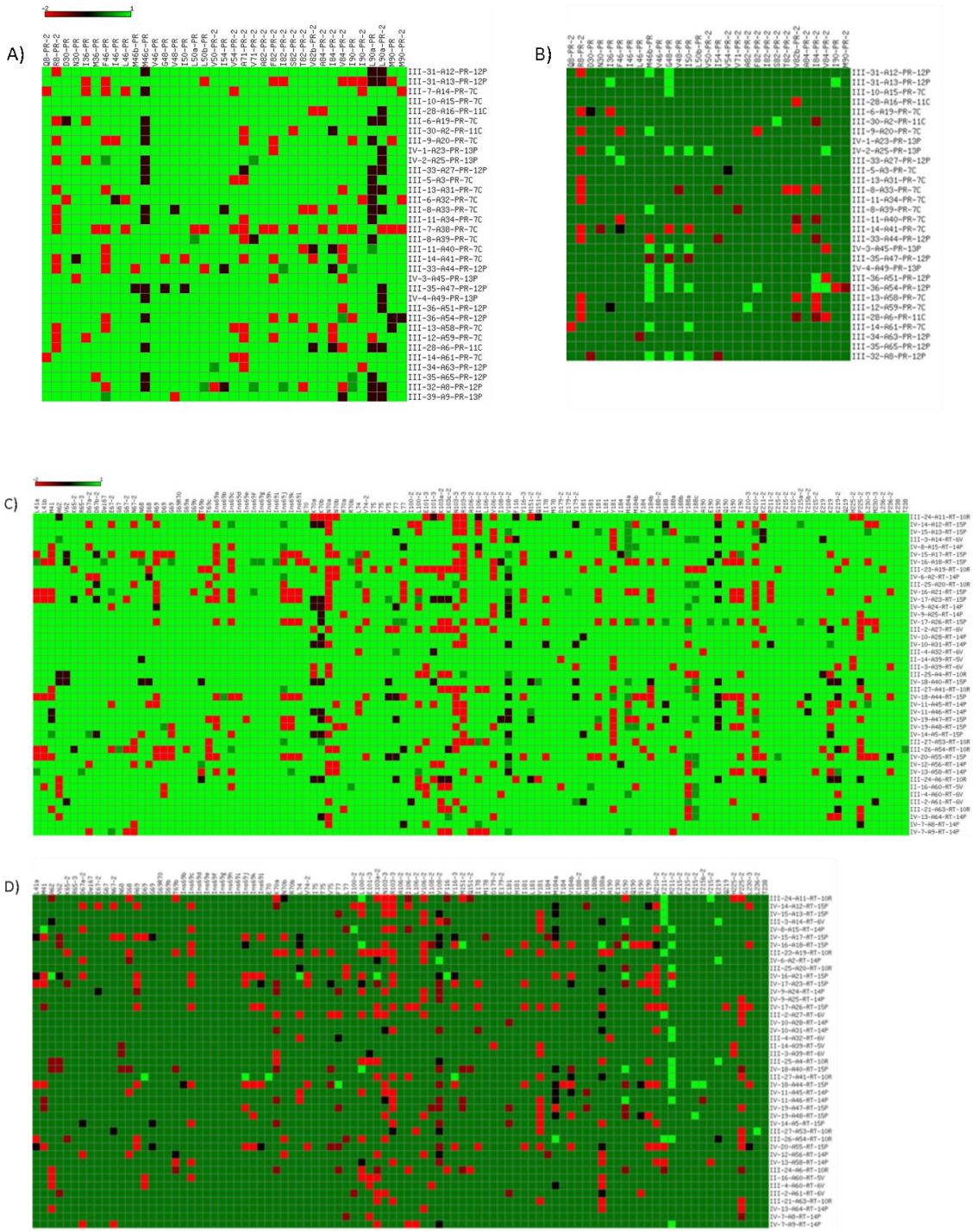
Supplementary Fig. S6.- Accumulation of errors (FP+FN+UD) during classification of clinical samples. A) Per probe (bar: 11 to 20 errors); **B)** Per sample (bar: 10 to 22 errors).



Supplementary Fig. S7.- Complete theoretical hybridization tables of binary mixtures of clonal samples. A) 1.95c9-2.94c64 and **B)** pWT-pINS. Legend: green, expected hybridization; red, not expected hybridization; black: partial hybridization (one mismatch between probe and target is allowed at the 5' or 3' nucleotide of the hybridizing sequence.



Supplementary Fig. S8.- Theoretical hybridization tables of clinical samples considering a detection limit of 10% for minority sequences. Only perfect probe-target hybridizations are allowed. Columns correspond to the queried codons included in probes belonging to the PR (A) or RT (B) region. Rows represent the consensus sequence of the quasispecies of each clinical sample. Legend: green, hybridization (>10% of the clonal sequences of the sample match with the probe); black, residual hybridization (matching in 5-10% of the clones); red, no hybridization.



Supplementary Fig. S9.- Classification accuracy of hybridized clinical samples. Columns, probes included in Virochip 4.0 corresponding to the PR (panels A and B) and RT (panels C and D) regions. Rows, hybridized samples. Panels A and C show the classification accuracy without filtering and Panels B and D show the effect of the stepwise filtering protocol applied to the hybridization data. Legend: dark green, correctly classified signal (TP or TN); dark red, FN signal; red, FP signal; black, UD spot; light green, no data (due to spots discarded during the quality control or to the absence of certain probes in different versions of the microarray).

Signals	Classification accuracy (% of signals) after filtering							
	Step 1: Spot filter		Step 2: Probe filter 1 (overlapped)		Step 3: Probe filter 2 (duplications)		Step 4: Array filter	
	1.95c9/ 2.94c64 (RT and PRRT)	pWT/ pINS (RT)	1.95c9/ 2.94c64 (RT and PRRT)	pWT/ pINS (RT)	1.95c9/ 2.94c64 (RT and PRRT)	pWT/ pINS (RT)	1.95c9/ 2.94c64 (RT and PRRT)	pWT/ pINS (RT)
Correct	95.32	91.15	97.80	93.08	97.64	94.26	97.64	94.26
FP	2.61	3.35	1.37	2.13	1.50	1.44	1.50	1.44
FN	1.44	4.77	0.49	3.93	0.54	3.36	0.54	3.36
UD	0.62	0.73	0.34	0.86	0.32	0.93	0.32	0.93

Supplementary Table S1.- **Classification accuracy of pure clonal samples used to set the sensitivity of detection in mixtures.** Legend: False positive (FP), False Negative (FN), Correct (TP +TN), Undefined (UD).

5.- Discusión

La presente tesis se ha llevado a cabo en el marco teórico de la bioinformática estructural, una ciencia multidisciplinar que nos sirve de ayuda y guía en la racionalización de los procesos moleculares de interés biológico, enfocada a predecir las interacciones entre proteínas identificadas como posibles dianas terapéuticas y potenciales fármacos. Con este objetivo, he utilizado e implementado métodos basados en las estructuras 3D de proteínas y pequeñas moléculas ya que, como hemos visto anteriormente, dada la estructura de una proteína podemos predecir el modo y energía de unión de posibles ligandos en su sitio activo con una fiabilidad cada vez mayor a medida que los métodos maduran.

La resistencia creciente que presentan ciertos microorganismos a los fármacos en uso es una de las motivaciones para el desarrollo de nuevos fármacos y el diseño de métodos que identifiquen las mutaciones involucradas en dicha resistencia para su aplicación en terapia personalizada. En esta tesis he estudiado dos casos donde nos enfrentamos a mecanismos de resistencia a fármacos por parte de la bacteria *Acinetobacter baumannii* (realizando una búsqueda de nuevos candidatos a través del cribado virtual de OXA-24) y del virus del VIH (presentado un protocolo computacional para la identificación de mutaciones e inserciones asociadas a la resistencia que presenta un paciente *a priori* para una correcta administración de los fármacos disponibles). El abordaje del diseño de fármacos basado en la técnica de *docking* presentado en esta tesis, parte de la necesidad de disponer de las estructuras 3D de proteínas y ligandos o de un conjunto representativo de sus conformaciones accesibles. Las estructuras pueden obtenerse mediante experimentos o mediante técnicas computacionales como el modelado por homología (HM) cuando no disponemos de ellas, siempre y cuando exista un molde adecuado con una relación evolutiva suficiente con la proteína diana. Por otra parte hemos visto que es importante no limitarse a una estructura estática, sino considerar la flexibilidad de la proteína, por ejemplo mediante técnicas de dinámica molecular (MD) y modos normales (NMA) que permiten mejorar el ajuste entre el potencial fármaco y el receptor durante el *docking*, evaluar la accesibilidad conformacional y la estabilidad de las moléculas involucradas o realizar la búsqueda de posibles sitios *drogables* de la proteína con el fin de delimitar el cálculo del *docking* a una región determinada, reduciendo el tiempo de cálculo y aumentando el porcentaje de éxito de los *hits* propuestos.

La investigación biomédica se ha visto beneficiada por la creciente disponibilidad de datos estructurales (Goodsell, Burley, & Berman, 2013b) que han permitido conocer a nivel atómico proteínas involucradas en procesos de salud y enfermedad, potenciando el descubrimiento de fármacos basado en su estructura. A pesar de que hablar de “la estructura” de una proteína sobreentiende una visión estática, como si una proteína fuera una sola

estructura del PDB, se va imponiendo, cada vez más una visión dinámica que hace hincapié en su variabilidad conformacional, extrema en el caso de las proteínas desordenadas que poseen largos fragmentos carentes de estructura definida y cuya relevancia biológica es cada vez más evidente (Uversky, 2013b). La variabilidad conformacional de las proteínas, que se traduce en la gran flexibilidad de su dinámica intrínseca, está modulada por modificaciones post-traduccionales, por sus interacciones con pequeños ligandos, otras proteínas y ácidos nucleicos, y se puede estudiar experimentalmente mediante técnicas de RMN u otras técnicas biofísicas. El problema de la heterogeneidad conformacional de las proteínas es difícil de abordar en el contexto del diseño de fármacos, ya que el *docking* suele estar basado, por razones de capacidad de cálculo, en la aproximación de proteína rígida, y por esto constituye uno de los mayores retos del diseño de fármacos basado en estructura.

La estructura y dinámica de las proteínas son características de gran interés dada su relación con la función que éstas desempeñan en los organismos. En décadas pasadas, se favorecía una visión estática de las proteínas (paradigma 1 secuencia -> 1 estructura -> 1 función). Sin embargo, en tiempos más recientes este paradigma ha sido sustituido por una visión más dinámica donde ni la secuencia es única (variabilidad de isoformas por *splicing*) ni lo es la estructura (la forma correcta de representar las estructuras de las proteínas es a través de su *ensemble* de equilibrio con el paisaje energético asociado, que a veces está caracterizado por distintos mínimos locales o sub-estados conformacionales entre los cuales la proteína puede moverse superando barreras energéticas) ni lo es la función (puede variar según las interacciones moleculares y las modificaciones pos-traduccionales). La variabilidad estructural es máxima en el caso de las proteínas desordenadas, cuya relevancia biológica es cada vez más evidente, y cuyos representantes estructurales pueden estar enormemente alejados entre sí, de manera que no se pueden resumir con ninguna estructura media estable (las proteínas desordenadas se han encontrado en esta tesis en el contexto del estudio del centrosoma humano). Asimismo, los fenómenos de alostería, que están íntimamente relacionados con la catálisis enzimática, refuerzan la idea de que el comportamiento dinámico de estas moléculas es necesario para su correcto funcionamiento (Nussinov & Tsai, 2013). Además, el control alostérico de la estructura y de la función de las proteínas puede surgir por modificaciones post-traduccionales, muy habituales en las células (por ejemplo la carboxilación de una lisina en el sitio activo en el caso de OXA-24).

He presentado un ejemplo de VS real en colaboración con el grupo del Dr. Antonio Romero, donde el 20% de los *hits* propuestos para la proteína OXA-24 actualmente se encuentran en fase de optimización tras las pruebas *in vivo*. Encontrar nuevos fármacos para

esta proteína es crucial ya que degrada los antibióticos administrados a pacientes con cuadros de resistencia. La disponibilidad de cristales de esta proteína puso en evidencia los residuos causantes de la resistencia que forman un túnel en el entorno del sitio activo reduciendo la accesibilidad a la cavidad y aumentando su especificidad. Conseguir superar la resistencia adquirida por las diferentes cepas bacterianas es una lucha constante que no deja tregua a la biomedicina. La importancia de OXA-24 y los esfuerzos por descifrar su funcionamiento se evidencia a través de cada nuevo cristal que se deposita en el PDB [4F94 (2013), 3PAE, 3PAG (2011) y 3MBZ (2011)] y las diferentes búsquedas de inhibidores (Pitout, 2013).

En un escenario ideal, las estructuras que la proteína y el ligando presentan en el complejo son conocidas. Si la estructura de la proteína se conoce pero la del ligando no, generamos los confórmeros de éste teniendo en cuenta reglas químicas de distancias y ángulos entre sus diferentes átomos así como la geometría en base al tipo de hibridación. La precisión de nuestro programa de *docking* depende de si somos capaces o no de generar la conformación que adopta el ligando en el sitio activo de la proteína. Cuando la conformación que el ligando adquiere en el complejo se encontraba entre las conformaciones de partida para el *docking* la precisión alcanzaba el 93% pero en caso contrario la precisión caía hasta el 55% (CDOCK con el conjunto de Astex y un límite de RMSD= 2 Å). En el programa de *docking* previo CDOCK tanto la fase de muestreo conformacional y como la fase de puntuación energética fue susceptible de mejoras. Por un lado requería de una mayor exploración del espacio conformacional el ligando y por otro lado la obtención de una puntuación más completa incluyendo nuevos términos en la función de *scoring*. Su implementación aumentó la precisión de CDOCK hasta el 61%. Después de las diferentes mejoras introducidas en esta tesis, la precisión obtenida con el nuevo programa CRDOCK (*Artículo 6*) alcanzó el 73% teniendo en cuenta la mejor pose en un *docking* con ligando flexible y proteína rígida, comparable con otros métodos como GLIDE, Autodock, etc. pero con un coste computacional menor lo que permite su implementación masiva en una plataforma de VS manteniendo tiempos de cálculo razonables. Sin embargo la caracterización de la flexibilidad de los sistemas de interés y su inclusión en los estudios de *docking* y VS puede enriquecer los resultados proporcionando candidatos más realistas, permitiendo además el abordaje de nuevas estrategias en el diseño de fármacos que consideren no solo sitios activos conocidos, sino sitios crípticos, *i.e.* aquellos sitios de unión no evidentes a partir de las estructuras estáticas resueltas, e incluso sitios alostéricos.

Como hemos visto a lo largo de la tesis, el resultado final del *docking* está fuertemente condicionado por la función de *scoring* o energía libre utilizada. Si las interacciones

moleculares relevantes no se encuentran correctamente definidas no es posible alcanzar la pose correcta. Se requiere de un nivel de detalle enorme (HB, desolvataciones, aguas cristalográficas, cargas atómicas correctas, etc.) aunque en algunos casos se suelen considerar funciones más simples que tienen en cuenta solo los términos *vdW* y coulómbico de la energía libre de unión. Habitualmente, en un protocolo de VS aumentamos el nivel de detalle molecular de la función de energía a medida que, aplicando filtros sucesivos, disminuye el número de ligandos potenciales a evaluar, evitando que los tiempos de cálculo se disparen. Tanto el estudio de las interacciones moleculares como la evaluación de las trayectorias de MD de los complejos predichos representan una fracción significativa del tiempo de ejecución de un proyecto de diseño de fármacos asistido por ordenador, por lo que cualquier aceleración en dicho proceso supone un avance, siempre y cuando la fiabilidad del método no se vea comprometida. Con el desarrollo de la función de scoring MM-ISMSA (*Artículo 5*) hemos creado una herramienta que reduce sustancialmente el tiempo de cálculo de energías libres en poses de *docking* y a lo largo de trayectorias de MD con resultados que se correlacionan bien con aquellos de otras herramientas más intensivas computacionalmente (por ejemplo MM-GBSA y MM-PBSA). El desarrollo de una interfaz gráfica dentro de *PyMOL* pretende extender su uso a usuarios no expertos permitiendo la inspección visual de los resultados de *docking* de un modo rápido y sencillo, no solo desde un punto de vista energético sino también gráfico a través de la distribución espacial de las interacciones más relevantes que se observan en un determinado complejo.

Dentro de las mejoras a nivel de la función de *scoring,* he incorporado un nuevo término de HB que ha sido determinante a la hora de obtener mejores soluciones de *docking* en los complejos del conjunto de prueba de Astex por su relevancia en la estabilización de interacciones moleculares, tanto es así que Lou *et. al* (Luo, Pei, & Zhu, 2010) propusieron un programa de *docking* guiado principalmente por geometría y la formación de HB. Sin embargo cabe destacar algunas consideraciones que pueden limitar el éxito del método implementado como es el protocolo empleado a la hora de añadir los átomos de hidrógeno. Dado que los átomos de hidrógeno no se incluyen en las estructuras de RX se deben añadir *a posteriori* teniendo en cuenta el estado de protonación de los residuos en base un pH dado (Word, Lovell, Richardson, & Richardson, 1999)(Li, Roy, & Zhang, 2009). Por ello, sería interesante repetir las pruebas llevadas a cabo en éste trabajo con un conjunto de complejos donde se incluyan las posiciones experimentales de los átomos de hidrógeno. Otra dificultad añadida son los átomos de hidrógeno que pueden rotar como en el grupo –OH de los alcoholes. En nuestro conjunto de datos algunos de los HB en los que este grupo se encontraban

involucrados no se identificaban porque los valores de sus variables geométricas se encontraban fuera de los rangos establecidos, lo que sería abordable si además de añadir los átomos de hidrógeno optimizáramos los valores de sus ángulos que pueden rotar libremente.

En el *Artículo 6* he presentado un método para refinar las soluciones del *docking* rígido a partir de conformaciones pre-generadas con el programa ALFA, un generador de confórmeros tridimensionales de ligandos desarrollado en nuestro laboratorio, usando nuestra nueva función de *scoring* MM-ISMSA. Con algunos ligandos resulta difícil generar la conformación adecuada, o bien por la complejidad de la molécula (*i.e.* número de enlaces rotables) o porque las conformaciones de los ligandos dentro de un complejo obtenido por RX pueden estar bastante influenciadas por la fuerza que ejerce la proteína sobre ellos y presentan desviaciones en ángulos de enlace y de torsión respecto a lo que supone nuestro algoritmo de construcción. Por esta razón, en esta tesis he desarrollado un método para minimizar, en el contexto del sitio activo, la energía libre de la pose del ligando predicha por el programa de *docking*, modificando sus enlaces rotables. Dicha modificación ha permitido mejorar en más de 1 Å el resultado del *docking* en el 20% de los ligandos estudiados. Como en el método que proponemos el ligando es totalmente flexible en el contexto del sitio activo, en principio el programa sería capaz de generar cualquier conformación permitida por las restricciones del sitio activo y la función de energía, incluso empezando por una única conformación del ligando. Este método se ha empleado de momento como una prueba de concepto para ver si los resultados de *docking* mejoraban a nivel de RMSD al minimizar la energía libre del complejo guiada por nuestra función de *scoring*. El mismo procedimiento se podría aplicar a las cadenas laterales de los residuos clave de la proteína en su interacción con el ligando, pero siempre teniendo en cuenta que un mayor número de ángulos torsionales aumentará el tiempo de computación. A pesar de lo prometedor del método presentado, para aumentar la capacidad predictiva del *docking* es necesario seguir trabajando en mejoras que contribuyan a la correcta orientación del ligando en el sitio activo, por ejemplo mediante la inclusión de la contribución por el apilamiento de anillos aromáticos, entropía etc. o la inclusión de elementos importantes del sistema como la flexibilidad de la proteína, las aguas cristalográficas, etc. habitualmente descartadas de los cálculos por simplicidad o por la ausencia de dicha información.

Uno de los grandes retos del diseño de fármacos es el que representan algunas dianas terapéuticas para las que no disponemos de estructuras determinadas experimentalmente, especialmente proteínas de membrana por sus dificultades añadidas a la hora de la purificación, cristalización, etc. El HM se presenta como un método valioso en la predicción de

estructuras de proteínas relacionadas evolutivamente. La robustez del HM se basa en la ID compartida entre las proteínas molde y diana, donde, cuanto mayor sea la ID más creíble será, en principio, el modelo 3D obtenido. Una de las limitaciones de la técnica de modelado comparativo es la disponibilidad de estructuras molde resueltas experimentalmente que presenten una ID respecto a la diana dentro de un rango de valores que permita su modelado con alta probabilidad de acierto, es decir, que se encuentre presente una ID mayor del 20-30%. En los últimos años el HM se ha visto respaldado por los proyectos de genómica estructural (Protein Structure Initiative, PSI) cuyo objetivo es la búsqueda y determinación estructural de plegamientos representativos de familias de proteínas mediante RX y NMR ampliando el rango de aplicabilidad del HM (Cavasotto & Phatak, 2009b).

En el *Artículo 1* presentamos un protocolo automático y modular de HM con un conjunto de parámetros modificables por el usuario y de aplicación a grandes conjuntos de proteínas de modo masivo sin tener que depender de servidores web como Swissprot, Protein model portal etc., permitiendo una mayor rapidez e independencia en la generación y posterior evaluación de los modelos obtenidos por homología. Entre los parámetros modificables destacamos: **(1)** el valor de corte de la ID para seleccionar el tipo de alineamiento (alineamiento de pares o alineamientos de perfiles HMM) permitiendo adaptar los tiempos de cálculo según la complejidad del caso; **(2)** la base de datos sobre la que realizar la búsqueda (bases de datos de perfiles HMM de todo el PDB incluyendo datos de RX y de RNM o reducidas para un tipo de plegamiento concreto o familias de proteínas disminuyendo el ruido en los alineamientos por proteínas de diferente plegamiento); **(3)** el molde a utilizar (el mejor molde disponible en base al número de residuos idénticos, todos los posibles moldes permitiendo estudiar la variabilidad estructural de una secuencia, o seleccionar un molde concreto de los propuestos explícitamente); **(4)** las regiones a considerar (solapantes o no) y **(5)** el número de posiciones contiguas sin molde debido a inserciones en la proteína diana permitidas en el alineamiento que serán generadas por técnicas *ab initio*.

El protocolo aplicado, determinado tratando de mantener un compromiso entre rapidez y calidad de los resultados, incluye la búsqueda mediante alineamiento por pares usando blast sobre una base de datos no redundante agrupada al 100% de ID (PDB100) cuando la ID molde-diana supera el 70%. En caso contrario, es decir para homólogos más lejanos, la búsqueda se lleva a cabo mediante alineamientos perfil-perfil de moldes en una base de datos de perfiles HMM generados a partir de las secuencias de las estructuras resueltas de PDB70. Con este objetivo he usado la herramienta HHblits (Remmert, Biegert, Hauser, & Söding, 2011) que permite la generación de perfiles HMM y su posterior búsqueda perfil-perfil en bases de

datos de un modo más rápido y eficaz respecto a su predecesor HHsearch. El considerar solo los aminoácidos que se asocian a estructuras resueltas experimentalmente en vez de las secuencias completas de las proteínas durante la fase de alineamiento perfil-perfil nos permite seleccionar los moldes con el mayor número de residuos alineados que pueden ser usados en la construcción 3D de la diana y no proteínas con regiones que no se pueden usar como moldes porque éstas son desordenadas o altamente flexibles y no se encuentran definidas experimentalmente. Como además buscamos maximizar la cobertura de la secuencia de la proteína diana, incorporamos la posibilidad de seleccionar los mejores moldes disponibles para cada una de las regiones modelables.

Un problema del modelado por homología es que, aún en presencia de una alta identidad de secuencia entre la proteína diana y la proteína molde, puede haber importantes regiones de *gaps* en el alineamiento debido a inserciones en la rama evolutiva que llega a la proteína diana o delecciones en la rama de la proteína molde. A la hora de modelar las regiones para las que no tenemos molde, sólo he construido *loops* de menos de 6 residuos para evitar construir modelos poco fiables. El proceso de modelado por homología a menudo produce estructuras que no son físicamente plausibles por tener colisiones entre átomos o enlaces químicos o HB con propiedades desfavorables desde el punto de vista energético. Para mejorar (refinar) las estructuras modeladas encontrando mejores representantes del *ensemble* de equilibrio de la proteína, se suelen someter los modelos con las mejores propiedades a un protocolo de dinámica molecular que elimine las distorsiones más evidentes. Sin embargo, para que este proceso de refinado se traduzca en una mejora no sólo de la energía, sino también del parecido entre el modelo y el *ensemble* de equilibrio de la proteína diana, dos puntos son muy importantes: **(1)** la dinámica tiene que tener lugar en una caja de agua (o en un entorno atómico que simule una membrana en el caso de proteínas de membranas), y **(2)** la simulación de la dinámica no tiene que ser demasiado larga, porque a veces la imprecisión de los campos de fuerza puede alejar demasiado a la estructura del modelo de la estructura del molde, reduciendo su parecido con la proteína diana. He programado un proceso de refinado automático, en el cual se seleccionan los modelos construidos por el programa MODELLER con las mejores puntuaciones del potencial estadístico DOPE, que penaliza contactos entre aminoácidos que se observan raramente y favorece los contactos frecuentes. Partiendo de estas estructuras, se aplica un refinado en dos pasos: **(1)** minimización en agua de la energía del campo de fuerzas de AMBER para eliminar colisiones atómicas y mejorar la geometría de los enlaces HB y enlaces químicos. Al terminar este proceso, la estructura alcanza un mínimo local de la energía en el cual al menos los choques estéricos han sido eliminados; y **(2)**

equilibrado elevando la temperatura del sistema hasta el valor deseado para alcanzar una estructura representativa del *ensemble* de equilibrio de la proteína a esta temperatura. Con estos dos procesos se obtienen pequeños reajustes en cadenas laterales que mejoran las distancias interatómicas y la red de HB inicial sin alterar demasiado su estructura.

El HM suele garantizar que tanto el plegamiento global (la disposición de elementos de estructuras secundaria) como la disposición espacial de los residuos del sitio activo presenten una disposición adecuada siempre que el alineamiento entre el molde y la diana sea correcto, porque estas propiedades son muy conservadas a lo largo de la evolución (sin embargo, el sitio activo puede cambiar si la proteína ha sufrido un cambio funcional y la función de la proteína molde y la proteína diana han divergido, por ejemplo modificando la especificidad enzimática). En esta situación "ideal" de conservación de la función, conservación del plegamiento y alineamiento correcto, las mayores diferencias entre el modelo y una estructura representativa de la proteína molde se encuentran en el empaquetamiento de las cadenas laterales y las conformaciones de los *loops* que se pueden mejorar con un proceso de refinado energético descrito arriba. Sin embargo, para obtener mejoras más sustanciales es necesario operar sobre los pasos iniciales del proceso de modelado **(1)** mejorar la información contenida en el espacio de secuencias consideradas en el alineamiento a través de la evaluación de su diversidad, la aplicación de filtros que reduzcan el número de falsos positivos, *i.e.* proteínas no relacionadas evolutivamente, y la edición iterativa del alineamiento final en zonas de ID baja o en regiones de grandes *loops* (corrigiendo en algún caso zonas con residuos alineados aislados dentro de ellos); **(2)** considerar matrices de sustitución específicas y diferentes según el tipo de proteína que modelemos (por ejemplo proteínas de membrana *vs.* globulares).

Durante esta tesis se ha estudiado no solo la estructura de proteínas sino también su desorden estructural. Gran parte de las proteínas de los organismos, en mayor o menor medida, presentan regiones intrínsecamente desordenadas en su estado nativo que son cruciales para su función reguladora y de señalización. Para estas proteínas el proceso de HM es poco útil porque la diversidad entre las estructuras representativas del *ensemble* de equilibrio es extrema. Sin embargo, se ha visto que muchas regiones desordenadas adquieren estructura cuando interaccionan con otra proteína o ácido nucleico. Se cree que esta propiedad de plegarse durante la interacción favorezca que las proteínas desordenadas tengan gran especificidad y baja afinidad, propiedades muy convenientes en las funciones de regulación molecular de los procesos celulares en las cuales las proteínas desordenadas están involucradas. En uno de los trabajos de investigación, llevado a cabo en colaboración con los grupos experimentales que formaban el consorcio Centrosome3D, caracterizamos

estructuralmente un conjunto de proteínas del centrosoma humano. El centrosoma humano, al igual que sus proteínas ortólogas en otras especies animales (Nido, Méndez, Pascual-García, Abia, & Bastolla, 2012, y datos no publicados de un estudio extendido a 36 especies), presenta una mayor fracción de desorden estructural en relación con el conjunto de proteínas control del mismo organismo. A su vez, la fracción de desorden predicha correlaciona con la complejidad de los organismos*, i.e.* el desorden estructural aumenta con el número de tipos celulares. A pesar del gran grado de desorden (57%), se pudo modelar por HM el 27,6% de los residuos del centrosoma. Cabe destacar que ciertas regiones fueron modeladas a pesar de ser predichas como desordenadas pudiendo estar asociadas a transiciones desorden–orden durante la unión a otras moléculas. En total el 23% de las proteínas centrosomales no se pudo construir por falta de moldes adecuados para ello.

Una vez que disponemos de una o más estructuras representativas de una proteína obtenidas tanto con métodos experimentales (ej. RX y RNM) como computacionales (ej. HM), es interesante investigar su flexibilidad y su dinámica intrínseca a una temperatura dada. Diversos estudios han demostrado el importante papel que juega la flexibilidad en la unión de ligandos y en la activación proteica, llevando a numerosos laboratorios a embarcarse en la predicción de los movimientos de las proteínas. Para este objetivo disponemos de dos grandes tipos de métodos: o bien la dinámica molecular, que construye el *ensemble* térmico de las proteínas simulando su dinámica con un campo de fuerza clásico y en un entorno representado de forma adecuada, o bien el método de los modos normales, que permite calcular de manera analítica todas las propiedades de la distribución de equilibrio conformacional de la proteína suponiendo que las conformaciones relevantes no se alejen mucho de la conformación de mínima energía donde la proteína estaría en equilibrio dinámico.

Los movimientos en el mundo microscópico están gobernados por las leyes de la mecánica cuántica (QM), a su vez gobernados por funciones de probabilidad, no por leyes deterministas. Del mismo modo, los enlaces químicos son nubes de electrones en movimiento, no algo fijo y direccional que se forma mecánicamente como establece la MD. Dada la complejidad de la mecánica cuántica, la MD se presenta como una alternativa razonable para reducir el tiempo de computación y permitir la simulación de sistemas con un número de átomos mucho mayor mediante la aplicación de aproximaciones basada en la física de Newton para simular los movimientos. La MD, una técnica en constante mejora a diferentes niveles, juega un papel cada vez más importante en el desarrollo de terapias farmacológicas aportando información acerca de eventos dinámicos tales como el reconocimiento molecular y la unión

de fármacos o de posibles candidatos a ellos, independientemente del modelo de unión propuesto (*induced fit*, *conformational selection*, etc.). La MD, a través de su muestreo conformacional, se puede aplicar en la identificación de sitios de unión alostéricos o reguladores y sitios crípticos se puede aplicar en mejorar la identificación de verdaderos ligandos respecto a *decoys* (señuelos) en protocolos de *docking* mediante la incorporación de la flexibilidad del receptor durante el proceso (usando múltiples conformaciones de entrada) o *a posteriori* (evaluando su estabilidad y fluctuaciones a lo largo del tiempo), y por último se puede aplicar para la estimación de la energía libre de unión o $\Delta G_{unión}$ del complejo proteína-ligando determinada solamente por la energía previa a la unión y la energía final, que nos permite estimar la afinidad con la que se unen los diferentes ligandos durante un VS o durante la optimización de los *hits* seleccionados. He aplicado con éxito la técnica de MD no solo en el refinado de estructuras de HM mejorando su puntuación por residuo en la mayoría de los casos, sino también en el análisis de la flexibilidad conformacional y posible mimetismo molecular entre diferentes péptidos unidos al complejo MHC de clase I en colaboración con el grupo del Dr. José Antonio López de Castro (CBMSO, *Artículo 3*), aportando información sobre el mecanismo compensatorio durante la unión a medida que aumenta la longitud de los péptidos en cuestión. La MD se ha aplicado además en la evaluación de la estabilidad de interacciones en complejos proteína-ligando provenientes de resultados de un VS obtenidos a partir de una conformación estática de la proteína diana. Durante un refinado corto (de 1-5 ns) los principales cambios están asociados a la adaptación de las estructuras al campo de fuerzas mediante pequeños ajustes en distancias, ángulos de enlace, etc. en su búsqueda del mínimo energético sin modificar el plegamiento global de la proteína como en el caso de modelos de HM y resultados de VS. Trayectorias más largas (de 10-50 ns) están enfocadas a conseguir un muestreo de las conformaciones accesibles de las moléculas que mantienen una energía libre favorable como las presentadas en el artículo 3. Durante la MD, cuando la duración el muestreo conformacional es insuficiente, corremos el riesgo de que las estructuras obtenidas obedezcan más a la estructura elegida como inicio de la simulación que al *ensemble* térmico de la proteína que se quiere estudiar. En particular, si la estructura inicial está separada del estado de equilibrio por una barrera de energía libre grande respecto a la energía térmica, es casi imposible que la simulación alcance el estado de equilibrio. Por lo tanto es posible que las propiedades del sistema cambien de forma relevante si alargamos los tiempos de cálculo por lo cual no debemos sacar conclusiones más allá del tiempo de simulación.

En esta tesis, he estudio las interacciones de varios péptidos (2 de un organismo patógeno y uno endógeno) con una de las proteínas del complejo MHC-I. Este sistema

presenta un gran interés porque se ha sugerido que en éste pueden estar involucrados fenómenos de mimetismo molecular, cuya relevancia en mecanismos patogénicos del sistema inmune está atrayendo un gran interés (Drayman et al., 2013). Teniendo en cuenta los resultados del péptido usado como control y otros resultados experimentales no publicados aún acerca del modo de unión de algunos de los péptidos utilizados en el estudio la conformación adoptada y el comportamiento dinámico diferencial de los péptidos en MD está en concordancia con los datos experimentales y bibliográficos hasta la fecha. Lo que aún queda por clarificar es la relevancia que pueda tener a nivel de su interacción con el TCR correspondiente validando o no el mimetismo molecular propuesto, tanto desde un punto de vista computacional como experimental debido a la dificultad para disponer de muestras de pacientes con los TCRs correspondientes con los que elaborar los experimentos y obtener su estructura.

En cuanto a la construcción de los complejos, cuando trabajamos con campos de fuerza, sea en MD o *docking*, el protocolo de preparación posee una relevancia especial. La protonación de los residuos en base al pKa de su micro-entorno, los *flips* de los residuos de Asn, Gln e His, la distribución alternativa de los átomos de hidrógenos asociada a diferentes átomos en estas últimas (HIE y HID) y el cálculo de las cargas atómicas parciales son algunos de los puntos críticos a revisar cuando los sistemas no se comportan como cabría esperar al compararlo con un control experimental simulado mediante el mismo protocolo que el sistema en estudio. Las principales limitaciones a las que se enfrenta la MD durante su fase de producción de la trayectoria son precisamente las aproximaciones asociadas al campo de fuerzas y su capacidad para realizar un muestreo conformacional exhaustivo en los tiempos de cálculos en los que nos movemos en la actualidad (habitualmente escala de varios nanosegundos a unos pocos microsegundos). Algunas de las limitaciones asociada al uso de un campo de fuerzas son: **(1)** la parametrización de residuos modificados, como la lisina carboxilada dentro del sitio activo de OXA-24, ligandos, cofactores y cualquier molécula no incluida en el campo de fuerzas, información requerida para el correcto funcionamiento de la MD y **(2)** las aproximaciones derivadas de la simplificación de los procesos microscópicos como es el caso de la formación y rotura de enlaces covalentes durante el VS de la proteína OXA-24 que une ligandos covalentemente al sitio activo. A pesar de ello, los complejos experimentales proteína-ligando unidos covalentemente, establecidos como control, se mantenían estables durante la MD aún sin considerar dicho enlace. Esto nos hace pensar que, aun así, podemos obtener resultados razonables para este tipo de sistemas. Para poder tener en cuenta los electrones y la formación/rotura de enlaces existen metodologías mixtas de mecánica cuántica

y mecánica molecular (QM/MM) que permiten seleccionar zonas de interés donde incrementar el nivel de detalle teórico, pero en todo caso está fuera de los objetivos de la presente tesis. Respecto al tiempo de cálculo, en numerosas publicaciones identificamos diversas medidas adoptadas para superar dicha limitación. Entre ellas caben destacar métodos computacionales a nivel de *software*, como la dinámica molecular acelerada, que reduce las barreras de energía entre conformaciones consiguiendo un muestreo conformacional más amplio en menos tiempo, y a nivel de *hardware* aprovechando la capacidad de cálculo que ofrecen las tarjetas gráficas (GPUs) en la aceleración de la producción de las trayectorias de MD, habitualmente en un orden de magnitud (Götz et al., 2012). Durante la realización de esta tesis hemos acelerado los cálculos de MD utilizando los módulos de cálculo en paralelo y en GPUs de los programas NAMD (MHC-I) y AMBER (refinado de soluciones de VS y modelos de HM del centrosoma).

Los cambios conformacionales normalmente ocurren en una escala de tiempos inaccesible por la MD, por lo que debemos considerar otras técnicas computacionales para su estudio. Recientemente, se ha hecho popular el estudio de los cambios de conformación a través del cálculo de los modos normales del modelo de red elástica, a pesar de que este cálculo solo tiene validez para desplazamientos muy pequeños desde una conformación de equilibrio dinámico porque las variaciones de energía siguen una aproximación armónica. Los modos normales son un conjunto de desplazamientos que permiten describir de la dinámica clásica de la molécula (como pequeñas fluctuaciones armónicas alrededor del punto de equilibrio) y su mecánica estadística (como superposición de pequeños desplazamientos independientes). Respecto a la MD, se obtienen resultados más robustos usando el modelo de red elástica, ya que no requiere de un campo de fuerza sino que lo construye, con un número muy pequeño de parámetros, a partir de la estructura representada en el PDB partiendo de la hipótesis que esta estructura corresponde a un mínimo del campo de fuerza y que el estado nativo es mínimamente frustrado, o sea, que todas las interacciones que se forman en este estado son energéticamente ventajosas. Las ventajas que presentan los NMA del modelo de red elástica (ENM) respecto a la MD es su simplicidad teórica, la velocidad de cálculo, la necesidad de pocos parámetros y su dependencia solo de la geometría, en este caso la topología de contactos, y la distribución de masas del sistema. En consecuencia, permite un muestreo mucho más exhaustivo en un tiempo reducido a costa de simplificar el detalle atómico a nivel de la contribución electrostática y el tipo de contacto establecido entre residuos, ambos importantes durante el reconocimiento molecular. A pesar de que la aproximación armónica en principio impediría que se puedan usar los modos normales para

estudiar grandes cambios de conformación, se ha observado que los cambios de conformación entre estructuras de la misma proteína que se encuentran en el PDB correlacionan muy fuertemente con los modos normales de baja frecuencia de los ENM, que representan movimientos colectivos de la proteína. En esta tesis se ha estudiado esta sorprendente relación, validando al mismo tiempo un modelo nulo de cambios de conformación y un nuevo modelo de red elástica en el espacio de los ángulos de torsión de la proteína, el Torsional Network Model [TNM,(Mendez & Bastolla, 2010)], ambos desarrollados en nuestro laboratorio. El TNM es un ENM que usa como grados de libertad los ángulos de torsión de la cadena principal. De esta manera, sólo hay dos grados de libertad por residuo en vez que los tres de los ENM Cartesianos (por ejemplo el ANM) que usan las coordenadas de los C_α. Como el número de grados de libertad influye fuertemente tanto en el coste computacional como en el consumo de memoria, el TNM es muy adecuado para estudiar la dinámica de equilibrio de sistemas de gran tamaño. Además, usar los ángulos de torsión permite representar de manera suficientemente precisa la dinámica de todos los átomos del esqueleto de la proteína y fijar longitud y ángulos de enlace a sus valores experimentales, impidiendo deformaciones que estarían muy penalizadas energéticamente.

En esta tesis he evaluado un modelo nulo de cambios de conformación, propuesto en el *Artículo 4,* que intenta explicar la observación de que las transiciones entre conformaciones diferentes de la misma proteína correlacionan con los modos normales predichos por los ENM. La explicación que se plantea propone un modelo nulo en el cual el cambio de conformación se debe a la respuesta lineal de la proteína, modelada como una red elástica, a una perturbación inducida por ejemplo por una unión molecular, por la fosforilación de un residuo, por un cambio en el medio (ej. pH) o por una mutación. Según el modelo nulo, la respuesta más probable a una perturbación genérica es una en la cual los desplazamientos producidos por el cambio de conformación se distribuyen a lo largo de los modos normales así como los desplazamientos producidos por la dinámica térmica, que son inversamente proporcionales a la frecuencia ω de cada modo normal α: $C_\alpha^2 \propto 1/\omega_\alpha^2$. De esta manera, se explica que los modos normales de baja frecuencia contribuyen más a los cambios de conformación porque las perturbaciones a lo largo de ellos hacen crecer menos la energía de la molécula.

He realizado un análisis masivo de todos los pares de estructuras en el PDB con la misma secuencia y con un cambio de conformación RMSD >1 Å que valida al mismo tiempo el modelo nulo y los modos normales predichos por el TNM. Para validar modelos de modos normales de redes elásticas, normalmente se siguen tres tipos de procedimientos: **(A)** Evaluación de la correlación entre los desplazamientos cuadráticos medios de cada átomo

predichos por NMA y aquellos que se infieren experimentalmente de la cristalografía de rayos X (*B-factors*). Sin embargo, los *B-factors* reflejan en gran medida movimientos de cuerpo rígido de la cadena proteica entera que no están representados por los modos normales, así que una baja correlación puede indicar simplemente que estos grados de libertad de cuerpo rígido tienen gran importancia en el cristal; **(B)** la comparación de los desplazamientos predichos por los modos normales y aquellos simulados con MD. Sin embargo, la validación de un método computacional con otro es una evaluación bastante débil; o **(C)** la comparación respecto a datos de RMN. La técnica de RNM es uno de los mejores métodos para evaluar la dinámica de equilibrio obtenida mediante NMA, ya que proporciona un conjunto de conformaciones representativas del *ensemble* de equilibrio. El problema es que no siempre disponemos de conjuntos de datos lo suficientemente representativos. El método que se presenta aquí constituye una alternativa interesante, ya que compara las predicciones de NMA que se obtienen a partir de una estructura del PDB con los cambios de conformación entre dos conformaciones determinadas en experimentos independientes. Los pares de estructuras del PDB se agruparon por condiciones experimentales (mismo/diferente cristal), en la forma unida o sin ligando a pesar de no considerarlos explícitamente (holo/apo), en la forma activa o inactiva de la proteína (residuos fosforilados o no) para tratar de obtener señales más claras. Observamos que, para la gran mayoría de los cambios de conformación observados, la contribución de cada modo normal a la dinámica térmica predicha por el TNM es proporcional a su contribución al cambio de conformación, lo que valida los modos normales predichos. Además, esta correlación es mucho más débil si se consideran cambios de conformación de menos de 1 Å de RMSD, caso en el cual el cambio de conformación está influenciado en buena medida por imprecisiones del experimento, lo cual confirma que la correlación observada no es un resultado trivial sino es un indicio de la calidad de los modos normales del TNM.

Aunque el modelo nulo describe bien la mayor parte de los cambios de conformación únicamente en base al principio físico de la respuesta de la proteína a una perturbación, en algunos casos observamos un gran número de pares con valores significativos de $\rho > 0$, lo que indica que los modos normales de baja frecuencia contribuyen a la dinámica térmica más de lo esperado según la respuesta lineal. En estos casos, a veces uno sólo de los modos normales de más baja frecuencia representa más del 70-80% del cambio de conformación. Estos casos son particularmente frecuentes para movimientos funcionales tales como reacciones enzimáticas y de transporte. Asimismo, encontré muchos valores de ρ significativos en conjuntos que representan cambios de conformación biológicamente relevantes, como los que siguen a la

fosforilación de un residuo o a la formación de un complejo de varias cadenas de proteínas (en particular, homopolímeros).

Cuando el valor de ρ es alto, el cambio de conformación se desarrolla a lo largo de los modos de baja frecuencia que producen un pequeño aumento de la energía a cambio de una deformación muy grande: un valor de ρ alto, por lo tanto, implica que el cambio de conformación tiene un coste energético menor. Por ello, hemos propuesto que el parámetro ρ se puede utilizar para identificar los cambios de conformación con posible valor funcional, a pesar de la dificultad de determinar *a priori* cuales son exactamente los modos normales de baja frecuencia que contribuyen a estos movimientos funcionales. De esta manera, la dinámica intrínseca de las proteínas descrita por los modos normales permite unificar los dos escenarios más importantes propuestos para la unión molecular: la selección conformacional y el ajuste inducido. En el modelo de la selección conformacional, las conformaciones en ausencia del ligando (apo) y en su presencia (holo) son accesibles a la proteína incluso cuando no hay ligando, cuyo papel es únicamente el de desplazar el equilibrio entre estas conformaciones. En el lenguaje de los modos normales, esta situación se corresponde cualitativamente con la existencia de una gran correlación entre la dinámica intrínseca de la proteína y la deformación producida por la unión del ligando ($\rho > 0$), o sea, el movimiento funcional existe aún en ausencia del ligando. Sin embargo, en el modelo del ajuste inducido la unión del ligando produce la perturbación que deforma la proteína. Este modelo se corresponde cualitativamente con lo que se espera bajo respuesta lineal, es decir el modelo nulo ($\rho \approx 0$). En resumen, el TNM se presenta como un método útil y rápido para el análisis de la dinámica intrínseca de las estructuras terciarias, el posible escenario del modo de unión y las similitudes alostéricas entre complejos, independientemente de la naturaleza de sus contactos.

Por otro lado, en esta tesis he trabajado en la mejora y aplicación de protocolos computacionales para el diseño de fármacos. Dichas mejoras y otras aportadas por otros grupos se están aplicando actualmente al estudio de los transportadores de glicina GlyT1 y GlyT2 en colaboración con el grupo de la Dra. Beatriz López Corcuera (CBMSO). Estos transportadores son importantes dianas terapéuticas en enfermedades que involucran sinapsis glicinérgicas y/o glutaminérgicas. Usando las estructuras cristalográficas disponibles de los transportadores de leucina (LeuT), un homólogo bacteriano lejano de GlyT1 y GlyT2 (ID alrededor del 20-25%, un valor bastante bajo), he modelado por homología tres conformaciones correspondientes a los pasos principales del transporte. Sin embargo, estas estructuras estáticas nos dan poca información acerca de las transiciones que tienen lugar durante el transporte. He decidido por lo tanto estudiar estas transiciones mediante el modelo

TNM, cuya aplicación a los cambios de conformación había validado con anterioridad, ampliando la información que tenemos de su flexibilidad y la accesibilidad conformacional, lo que nos permite generar un conjunto representativo de estructuras con el sitio activo abierto para poder realizar el *docking* de inhibidores conocidos y una búsqueda por VS de nuevos *hits* en sitios alternativos. El desarrollo de un nuevo campo de fuerzas de lípidos para AMBER, LIPID11 (Skjevik, Madej, Walker, & Teigen, 2012b), me ha permitido evaluar la estabilidad de los complejos mediante simulaciones de MD más realistas al incluir la bicapa lipídica. Estamos estudiando, además, la influencia de las mutaciones conocidas en la dinámica de las proteínas y en la unión a los sustratos e inhibidores.

Por último, hemos estudiado la variabilidad molecular del proceso de infección por HIV para abordar la emergencia de resistencias a fármacos (*Artículo 7*). Tan importante como disponer de un conjunto creciente de fármacos es el ser capaz de seleccionar para cada paciente cual es el más adecuado entre todos los disponibles, lo que llamamos terapia personalizada. En el caso del virus del VIH, la extrema variabilidad genética del virus, debida a su alta tasa de mutación, permite que algunos de sus clones adquieran resistencia a diversidad de fármacos dependiendo de las mutaciones o inserciones adquiridas. Para luchar contra este fenómeno, he contribuido a desarrollar un programa que permite caracterizar de un modo rápido y fiable una larga lista de variantes que presentan mutaciones y/o inserciones que confieren resistencia a los fármacos antiretrovirales que inhiben a las proteínas PR o RT. Este trabajo se ha desarrollado en colaboración con los grupos del Dr. Esteban Domingo (CBMSO) y del Dr. Carlos Briones (CAB).

A pesar de los continuos avances en el campo de los *microarrays*, la generalización del análisis de datos sigue siendo un factor limitante para su rápida aplicabilidad ya que en muchos casos se requieren de nuevas implementaciones dependiendo de la cuestión que queramos estudiar. Otra gran limitación está asociada a la variabilidad a varios niveles: desde el protocolo experimental (plataformas de uno o dos canales, la longitud de las sondas, su distribución espacial, su temperatura de hibridación óptima, el tamaño de muestra, las condiciones de laboratorio, etc.) al protocolo computacional (método de análisis de datos, filtrado de datos y valores de corte asociados, criterios de control de calidad, niveles de detección, etc.). Para tratar de reducir parte de la variabilidad experimental realizamos el pre-procesado de los datos mediante una normalización que permite distinguir más claramente hibridaciones positivas y negativas. Con ello conseguimos clasificar correctamente entre un 84% y un 93% (dependiendo del conjunto de datos). Durante el control de calidad la inclusión de los diferentes filtros mejoró la clasificación al reducir, principalmente, los falsos positivos

(*i.e.* aquellos casos con señal de fluorescencia debido a hibridaciones no específicas o contaminación). Con ello mejoramos en varios puntos el porcentaje de *spots* clasificados correctamente (entre un 89% y casi un 98% dependiendo del conjunto de datos). A pesar de los filtros, ciertas sondas y *microarrays* acumulan aún un gran número de errores de clasificación. Esto puede ser debido a los parámetros elegidos para el control de calidad automático, como el porcentaje de solapamiento que se permite entre la señal positiva y la señal negativa, o debido a las condiciones experimentales. Además, las señales positiva y negativa de algunas de las sondas no se pudieron caracterizar por ausencia de alguna de las mutaciones en las muestras hibridadas. Aun cuando no disponemos de datos de la señal que tendría una sonda determinada contamos con curvas de caracterización globales (*i.e*. generadas a partir de todo el conjunto de datos de entrenamiento) para su clasificación, sin embargo perdemos el detalle del comportamiento diferencial de las sondas (temperatura óptima de hibridación, señal, etc.) lo que puede afectar a su clasificación final.

Otro de los aspectos relevantes del estudio ha sido el determinar el umbral de detección de esta técnica, *i.e.* en qué proporción debe estar presente un clon con una determinada mutación para que sea detectable dentro de la población presente en un paciente. Aunque las mezclas de dos clones analizadas presentaban niveles de detección variables, en la mayoría de los casos un 10% es suficiente. Uno de los objetivos de este trabajo fue el de detectar no sólo las variantes mayoritarias sino aquellas minoritarias, porque esta información ayuda a predecir la evolución de la población vírica a lo largo del tiempo. Cuando atacamos las variantes mayoritarias con un fármaco, éstas disminuyen permitiendo que alguna de las minoritarias las sustituyan por lo que el tratamiento no solo debe ser personalizado sino probablemente también variable a lo largo del tiempo.

6.- Conclusiones

En esta tesis se han presentado diferentes protocolos *in silico* enfocados a mejorar la caracterización estructural, dinámica y de las interacciones entre proteínas y ligandos así como su aplicación al desarrollo de nuevos fármacos frente a dianas terapéuticas.

Como se ha evidenciado en los 7 artículos presentados, el trabajo que he llevado a cabo ha permitido:

1.- Mejorar las predicciones teóricas de las interacciones entre proteínas y ligandos durante el *docking* y el VS. Por un lado con la incorporación de un término para los HB en la función de puntuación de MM-ISMSA y por otro con la inclusión de los grados de libertad torsionales durante la minimización de las poses de los ligandos dentro del programa de *docking* CRDOCK.

2.- Buscar nuevos fármacos mediante VS de la proteína bacteriana OXA-24 responsable de resistencia a antibióticos, obteniendo candidatos prometedores en fase de optimización.

3.- Estudiar, mediante técnicas de modelado y dinámica molecular, la flexibilidad y la variabilidad conformacional de diversos péptidos (endógenos y patógenos) en su unión a la proteína HLA-B27*05 del complejo MHC de clase I. Esto nos ha permitido evaluar computacionalmente el posible mimetismo molecular entre el péptido endógeno y los patógenos pudiendo favorecer la cronicidad de la artritis reactiva.

4.- Desarrollar un protocolo automático de modelado por homología y aplicarlo al estudio de proteínas centrosomales humanas y a secuencias derivadas de simulaciones de evolución de proteínas. Los modelos 3D de las proteínas centrosomales obtenidos, así como otros datos de interés, se han puesto a disposición de la comunidad científica en una base de datos *on-line*.

5.- Validar un modelo nulo de los cambios conformacionales de las proteínas basado en modos normales torsionales que explica la correlación observada entre los modos normales de baja frecuencia y los cambios conformacionales. Dentro de este marco, proponemos que los cambios de conformación que se alejan significativamente del modelo nulo tienen relevancia en la función de la proteína.

6.- Diseñar e implementar la parte computacional de un protocolo para la identificación de clones de VIH que presentan resistencia a antiretrovirales, con el objetivo de permitir el desarrollo de terapias personalizadas. El protocolo incluye la cuantificación y el control de calidad de señales de fluorescencia provenientes de *microarrays* de expresión.

7.- Bibliografía

Abraham, M. H., Ibrahim, A., Zissimos, A. M., Zhao, Y. H., Comer, J., & Reynolds, D. P. (2002). Application of hydrogen bonding calculations in property based drug design. *Drug discovery today*, *7*(20), 1056–63.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403–10. doi:10.1016/S0022-2836(05)80360-2

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389–402.

Alvarez-Navarro, C., Cragnolini, J. J., **Dos Santos, H. G**., Barnea, E., Admon, A., Morreale, A., & López de Castro, J. A. (2013). Novel HLA-B27-restricted epitopes from Chlamydia trachomatis generated upon endogenous processing of bacterial proteins suggest a role of molecular mimicry in reactive arthritis. *The Journal of biological chemistry*. doi:10.1074/jbc.M113.493247

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science (New York, N.Y.)*, *181*(4096), 223–30.

Arenas, M., **Dos Santos, H. G.,** Posada, D., & Bastolla, U. (2013). Protein Evolution along Phylogenetic Histories under Structurally Constrained Substitution Models. *Bioinformatics*, btt530–. doi:10.1093/bioinformatics/btt530

Arnold, K., Bordoli, L., Kopp, J., & Schwede, T. (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics (Oxford, England)*, *22*(2), 195–201. doi:10.1093/bioinformatics/bti770

Arnold, K., Kiefer, F., Kopp, J., Battey, J. N. D., Podvinec, M., Westbrook, J. D., … Schwede, T. (2009). The Protein Model Portal. *Journal of structural and functional genomics*, *10*(1), 1–8. doi:10.1007/s10969-008-9048-5

Bahar, I., Lezon, T. R., Yang, L.-W., & Eyal, E. (2010). Global dynamics of proteins: bridging between structure and function. *Annual review of biophysics*, *39*, 23–42. doi:10.1146/annurev.biophys.093008.131258

Baker, D. (2000). A surprising simplicity to protein folding. *Nature*, *405*(6782), 39–42. doi:10.1038/35011000

Buchan, D. W. A., Ward, S. M., Lobley, A. E., Nugent, T. C. O., Bryson, K., & Jones, D. T. (2010). Protein annotation and modelling servers at University College London. *Nucleic acids research*, *38*(Web Server issue), W563–8. doi:10.1093/nar/gkq427

Burley, S. K., & Petsko, G. A. (1985). Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science (New York, N.Y.)*, *229*(4708), 23–8.

Cabrera, Á. C., Gil-Redondo, R., Perona, A., Gago, F., & Morreale, A. (2011). VSDMIP 1.5: an automated structure- and ligand-based virtual screening platform with a PyMOL graphical user interface. *Journal of computer-aided molecular design*, *25*(9), 813–24. doi:10.1007/s10822-011-9465-6

Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., … Woods, R. J. (2005). The Amber biomolecular simulation programs. *Journal of computational chemistry*, *26*(16), 1668–88. doi:10.1002/jcc.20290

Cavasotto, C. N., & Phatak, S. S. (2009). Homology modeling in drug discovery: current trends and applications. *Drug Discovery Today*, *14*(13), 676–683.

Chandonia, J.-M., & Brenner, S. E. (2006). The impact of structural genomics: expectations and outcomes. *Science (New York, N.Y.)*, *311*(5759), 347–51. doi:10.1126/science.1121018

Chen, K., & Kurgan, L. (2009). Investigation of atomic level patterns in protein--small ligand interactions. *PloS one*, *4*(2), e4473. doi:10.1371/journal.pone.0004473

Chiu, C. Y., Urisman, A., Greenhow, T. L., Rouskin, S., Yagi, S., Schnurr, D., … Ganem, D. (2008). Utility of DNA microarrays for detection of viruses in acute respiratory tract infections in children. *The Journal of pediatrics*, *153*(1), 76–83. doi:10.1016/j.jpeds.2007.12.035

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, *5*(4), 823–826.

Cortés Cabrera, Á., Klett, J., **Dos Santos, H. G.,** Perona, A., Gil-Redondo, R., Francis, S. M., … Morreale, A. (2012). CRDOCK: an ultrafast multipurpose protein-ligand docking tool. *Journal of chemical information and modeling*, *52*(8), 2300–2309. doi:10.1021/ci300194a

Deane, C. M., & Blundell, T. L. (2001). CODA: a combined algorithm for predicting the structurally variable regions of protein models. *Protein science : a publication of the Protein Society*, *10*(3), 599–612. doi:10.1110/ps.37601

DiMasi, J. A., Hansen, R. W., & Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, *22*(2), 151–185.

Dokholyan, N. V, Shakhnovich, B., & Shakhnovich, E. I. (2002). Expanding protein universe and its origin from the biological Big Bang. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(22), 14132–6. doi:10.1073/pnas.202497999

Dos Santos, H. G., Abia, D., Janowski, R., Mortuza, G., Bertero, M. G., Boutin, M., … Serrano, L. (2013). Structure and non-structure of centrosomal proteins. *PloS one*, *8*(5), e62633. doi:10.1371/journal.pone.0062633

Drayman, N., Glick, Y., Ben-nun-shaul, O., Zer, H., Zlotnick, A., Gerber, D., … Oppenheim, A. (2013). Pathogens use structural mimicry of native host ligands as a mechanism for host receptor engagement. *Cell host & microbe*, *14*(1), 63–73. doi:10.1016/j.chom.2013.05.005

Dymock, B. W., Barril, X., Brough, P. A., Cansfield, J. E., Massey, A., McDonald, E., … Drysdale, M. J. (2005). Novel, potent small-molecule inhibitors of the molecular chaperone Hsp90 discovered through structure-based design. *Journal of medicinal chemistry*, *48*(13), 4212–5. doi:10.1021/jm050355z

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics (Oxford, England)*, *14*(9), 755–63.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*(5), 1792–7. doi:10.1093/nar/gkh340

Fan, H., Irwin, J. J., & Sali, A. (2012). Virtual Ligand Screening Against Comparative Protein Structure Models. *Methods in Molecular Biology (Clifton, N.J.)*, *819*, 105–126. doi:10.1007/978-1-61779-465-0_8

Fidelis, K., Stern, P. S., Bacon, D., & Moult, J. (1994). Comparison of systematic search and database methods for constructing segments of protein structure. *Protein engineering*, *7*(8), 953–60.

Fiser, A., & Šali, A. (2003). Modeller: Generation and Refinement of Homology-Based Protein Structure Models. In J. and R. M. S. Charles W. Carter (Ed.), (Vol. Volume 374, pp. 461–491). Academic Press.

Fogolari, F., Brigo, A., & Molinari, H. (2002). The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *Journal of molecular recognition : JMR*, *15*(6), 377–92. doi:10.1002/jmr.577

Goodsell, D. S., Burley, S. K., & Berman, H. M. (2013). Revealing structural views of biology. *Biopolymers*, *99*(11), 817–824. doi:10.1002/bip.22338

Götz, A. W., Williamson, M. J., Xu, D., Poole, D., Le Grand, S., & Walker, R. C. (2012). Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *Journal of Chemical Theory and Computation*, *8*(5), 1542–1555. doi:10.1021/ct200909j

Gullingsrud, J., & Schulten, K. (2004). Lipid bilayer pressure profiles and mechanosensitive channel gating. *Biophysical journal*, *86*(6), 3496–509. doi:10.1529/biophysj.103.034322

Hartshorn, M. J., Verdonk, M. L., Chessari, G., Brewerton, S. C., Mooij, W. T. M., Mortenson, P. N., & Murray, C. W. (2007). Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of medicinal chemistry*, *50*(4), 726–41. doi:10.1021/jm061277y

Irwin, J. J., & Shoichet, B. K. (2005). ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model*, *45*(1), 177–182.

Jo, S., Lim, J. B., Klauda, J. B., & Im, W. (2009). CHARMM-GUI Membrane Builder for mixed bilayers and its application to yeast membranes. *Biophysical journal*, *97*(1), 50–58. doi:10.1016/j.bpj.2009.04.013

Jones, D. T., & Ward, J. J. (2003). Prediction of disordered regions in proteins from position specific score matrices. *Proteins*, *53 Suppl 6*, 573–8. doi:10.1002/prot.10528

Jorgensen, W. L. (2004). The many roles of computation in drug discovery. *Science*, *303*(5665), 1813–1818. doi:10.1126/science.1096361 303/5665/1813 [pii]

Jurecka, P., Sponer, J., Cerný, J., & Hobza, P. (2006). Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Physical chemistry chemical physics : PCCP*, *8*(17), 1985–93. doi:10.1039/b600027d

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, *32*(5), 922–923. doi:10.1107/S0567739476001873

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–637. doi:10.1002/bip.360221211

Katoh, K., & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics*, *9*(4), 286–98. doi:10.1093/bib/bbn013

Kellenberger, E., Rodrigo, J., Muller, P., & Rognan, D. (2004). Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins*, *57*(2), 225–42. doi:10.1002/prot.20149

Klett, J., Núñez-Salgado, A., **Dos Santos, H. G.,** Cortés-Cabrera, Á., Perona, A., Gil-Redondo, R., … Morreale, A. (2012). MM-ISMSA: An Ultrafast and Accurate Scoring Function for Protein–Protein Docking. *Journal of Chemical Theory and Computation*, *8*(9), 3395–3408. doi:10.1021/ct300497z

Krogh, A., Brown, M., Mian, I. S., Sjölander, K., & Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. *Journal of molecular biology*, *235*(5), 1501–31. doi:10.1006/jmbi.1994.1104

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology*, *305*(3), 567–80. doi:10.1006/jmbi.2000.4315

Laskowski, R., Macarthur, M., Moss, D., & Thornton, J. (1993). {PROCHECK}: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, *26*, 283 – 291.

Leopold, J. L., & Frank, R. L. (2012). Protein secondary structure prediction using BLAST and exhaustive RT-RICO, the search for optimal segment length and threshold. In *2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (pp. 35–42). IEEE. doi:10.1109/CIBCB.2012.6217208

Li, Y., Roy, A., & Zhang, Y. (2009). HAAD: A quick algorithm for accurate prediction of hydrogen atoms in protein structures. (A. Hofmann, Ed.)*PloS one*, *4*(8), e6701. doi:10.1371/journal.pone.0006701

Liberles, D. A., Teichmann, S. A., Bahar, I., Bastolla, U., Bloom, J., Bornberg-Bauer, E., … Whelan, S. (2012). The interface of protein structure, protein biophysics, and molecular evolution. *Protein science : a publication of the Protein Society*, *21*(6), 769–85. doi:10.1002/pro.2071

Lindahl, E., & Sansom, M. S. (2008). Membrane proteins: molecular dynamics simulations. *Current Opinion in Structural Biology*, *18*(4), 425–431. doi:10.1016/j.sbi.2008.02.003

Liu, J., Zheng, Q., Deng, Y., Cheng, C.-S., Kallenbach, N. R., & Lu, M. (2006). A seven-helix coiled coil. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(42), 15457–62. doi:10.1073/pnas.0604871103

Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G., & Chothia, C. (2000). SCOP: a structural classification of proteins database. *Nucleic acids research*, *28*(1), 257–9.

Luo, W., Pei, J., & Zhu, Y. (2010). A fast protein-ligand docking algorithm based on hydrogen bond matching and surface shape complementarity. *Journal of molecular modeling*, *16*(5), 903–13. doi:10.1007/s00894-009-0598-7

Lupas, A., Van Dyke, M., & Stock, J. (1991). Predicting coiled coils from protein sequences. *Science (New York, N.Y.)*, *252*(5009), 1162–4. doi:10.1126/science.252.5009.1162

Mendez, R., & Bastolla, U. (2010). Torsional network model: normal modes in torsion angle space better correlate with conformation changes in proteins. *Physical Review Letters*, *104*(22), 228103.

Morreale, A., Gil-Redondo, R., & Ortiz, A. R. (2007). A new implicit solvent model for protein-ligand docking. *Proteins*, *67*(3), 606–616.

Mueckler, M., & Makepeace, C. (2008). Transmembrane segment 6 of the Glut1 glucose transporter is an outer helix and contains amino acid side chains essential for transport activity. *The Journal of biological chemistry*, *283*(17), 11550–5. doi:10.1074/jbc.M708896200

Myers, S., & Baker, A. (2001). Drug discovery--an operating model for a new era. *Nature biotechnology*, *19*(8), 727–30. doi:10.1038/90765

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453.

Nido, G. S., Méndez, R., Pascual-García, A., Abia, D., & Bastolla, U. (2012). Protein disorder in the centrosome correlates with complexity in cell types number. *Molecular bioSystems*, *8*(1), 353–67. doi:10.1039/c1mb05199g

Nogales-Cadenas, R., Abascal, F., Díez-Pérez, J., Carazo, J. M., & Pascual-Montano, A. (2009). CentrosomeDB: a human centrosomal proteins database. *Nucleic acids research*, *37*(Database issue), D175–80. doi:10.1093/nar/gkn815

Nussinov, R., & Tsai, C.-J. (2013). Allostery in disease and in drug discovery. *Cell*, *153*(2), 293–305. doi:10.1016/j.cell.2013.03.034

Onufriev, A., Case, D. A., & Bashford, D. (2002). Effective Born radii in the generalized Born approximation: the importance of being perfect. *Journal of computational chemistry*, *23*(14), 1297–304. doi:10.1002/jcc.10126

Orengo, C. A., Pearl, F. M., Bray, J. E., Todd, A. E., Martin, A. C., Lo Conte, L., & Thornton, J. M. (1999). The CATH Database provides insights into protein structure/function relationships. *Nucleic acids research*, *27*(1), 275–9.

Pascual-García, A., Abia, D., Ortiz, A. R., & Bastolla, U. (2009). Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures. *PLoS Computational Biology*, *5*(3), e1000331. doi:10.1371/journal.pcbi.1000331

Pérez, C., & Ortiz, A. R. (2001). Evaluation of Docking Functions for Protein−Ligand Docking. *Journal of Medicinal Chemistry*, *44*(23), 3768–3785. doi:10.1021/jm010141r

Pieper, U., Webb, B. M., Barkan, D. T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., … Sali, A. (2011). ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic acids research*, *39*(Database issue), D465–74. doi:10.1093/nar/gkq1091

Pitout, J. D. D. (2013). Enterobacteriaceae that produce extended-spectrum β-lactamases and AmpC β-lactamases in the community: the tip of the iceberg? *Current pharmaceutical design*, *19*(2), 257–63.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., … Finn, R. D. (2012). The Pfam protein families database. *Nucleic acids research*, *40*(Database issue), D290–301. doi:10.1093/nar/gkr1065

Quackenbush, J. (2001). Computational analysis of microarray data. *Nature Reviews Genetics*, *2*(6), 418–427. doi:10.1038/35076576

Rapp, C. S., & Friesner, R. A. (1999). Prediction of loop geometries using a generalized born model of solvation effects. *Proteins*, *35*(2), 173–83.

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, *9*(2), 173–175. doi:10.1038/nmeth.1818

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods*, *9*(2), 173–5. doi:10.1038/nmeth.1818

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein engineering*, *12*(2), 85–94.

Ruddy, K. J., Desantis, S. D., Gelman, R. S., Wu, A. H. B., Punglia, R. S., Mayer, E. L., … Burstein, H. J. (2013). Personalized medicine in breast cancer: tamoxifen, endoxifen, and CYP2D6 in clinical practice. *Breast cancer research and treatment*. doi:10.1007/s10549-013-2700-1

Sadreyev, R. I., & Grishin, N. V. (2006). Exploring dynamics of protein structure determination and homology-based prediction to estimate the number of superfamilies and folds. *BMC structural biology*, *6*(1), 6. doi:10.1186/1472-6807-6-6

Sanchez-Ruiz, J. M. (1995). Differential scanning calorimetry of proteins. *Sub-cellular biochemistry*, *24*, 133–76.

Santillana, E., Beceiro, A., Bou, G., & Romero, A. (2007). Crystal structure of the carbapenemase OXA-24 reveals insights into the mechanism of carbapenem hydrolysis. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(13), 5354–9. doi:10.1073/pnas.0607557104

Santos, H. G. Dos, Klett, J., Méndez, R. & Bastolla, U.(2013). Characterizing conformation changes in proteins through the torsional elastic response. *Biochimica et biophysica acta*, *1834*(5), 836–46. doi:10.1016/j.bbapap.2013.02.010

Sevier, C. S., & Kaiser, C. A. (2002). Formation and transfer of disulphide bonds in living cells, *3*(11), 836–847.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of molecular biology*, *213*(4), 859–83.

Skjevik, Å. A., Madej, B. D., Walker, R. C., & Teigen, K. (2012). LIPID11: a modular framework for lipid simulations using amber. *The journal of physical chemistry. B*, *116*(36), 11124–11136. doi:10.1021/jp3059992

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195–197.

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics (Oxford, England)*, *21*(7), 951–60. doi:10.1093/bioinformatics/bti125

Song, C. M., Lim, S. J., & Tong, J. C. (2009). Recent advances in computer-aided drug design. *Briefings in bioinformatics*, *10*(5), 579–91. doi:10.1093/bib/bbp023

Steiner, T. (2002). The hydrogen bond in the solid state. *Angewandte Chemie (International ed. in English)*, *41*(1), 49–76.

Stewart, J. J. P. (1990). MOPAC: A semiempirical molecular orbital program. *Journal of Computer-Aided Molecular Design*, *4*(1), 1–103. doi:10.1007/BF00128336

Studer, R. A., Dessailly, B. H., & Orengo, C. A. (2013). Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *The Biochemical journal*, *449*(3), 581–594. doi:10.1042/BJ20121221

Sun, W., Gerth, C., Maeda, A., Lodowski, D. T., Van Der Kraak, L., Saperstein, D. A., … Palczewski, K. (2007). Novel RDH12 mutations associated with Leber congenital amaurosis and cone-rod dystrophy: biochemical and clinical evaluations. *Vision research*, *47*(15), 2055–66. doi:10.1016/j.visres.2007.04.005

Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., & Abola, E. E. (1998). Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta crystallographica. Section D, Biological crystallography*, *54*(Pt 6 Pt 1), 1078–84.

Taverna, D. M., & Goldstein, R. A. (2002). Why are proteins marginally stable? *Proteins*, *46*(1), 105–9.

Tirion, M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical review letters*, *77*(9), 1905–1908.

Trevino, V., Falciani, F., & Barrera-Saldana, H. A. (2007). DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research. *Molecular Medicine*, *13*(9-10), 527–541. doi:10.2119/2006-00107.Trevino

Update on activities at the Universal Protein Resource (UniProt) in 2013. (2013). *Nucleic acids research*, *41*(Database issue), D43–7. doi:10.1093/nar/gks1068

Uversky, V. N. (2013). A decade and a half of protein intrinsic disorder: biology still waits for physics. *Protein science: a publication of the Protein Society*, *22*(6), 693–724. doi:10.1002/pro.2261

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., & Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins*, *52*(4), 609–623. doi:10.1002/prot.10465

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., & Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of molecular biology*, *337*(3), 635–645. doi:10.1016/j.jmb.2004.02.002

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, *28*(1), 31–36. doi:10.1021/ci00057a005

Word, J. M., Lovell, S. C., Richardson, J. S., & Richardson, D. C. (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of molecular biology*, *285*(4), 1735–47. doi:10.1006/jmbi.1998.2401

Xiang, Z. (2006). Advances in homology protein structure modeling. *Current protein & peptide science*, *7*(3), 217–227.

Xu, J., Jiao, F., & Yu, L. (2008). Protein structure prediction using threading. *Methods in molecular biology (Clifton, N.J.)*, *413*, 91–121.

Ying, L., & Sarwal, M. (2009). In praise of arrays. *Pediatric nephrology (Berlin, Germany)*, *24*(9), 1643–1659; quiz 1655, 1659. doi:10.1007/s00467-008-0808-z